Essays in Experimental Economics

Preferences and Information

Hedegaard, Morten

Publication date:
2011

Document version
Early version, also known as pre-print

Citation for published version (APA):
Hedegaard, M. (2011). *Essays in Experimental Economics: Preferences and Information*. Department of Economics, University of Copenhagen. PhD Series Vol. 2011 No. 146

FACULTY OF SOCIAL SCIENCES
UNIVERSITY OF COPENHAGEN

PhD Thesis

Morten Hedegaard

Essays in Experimental Economics: Preferences and Information

Academic advisor: Jean-Robert Tyran

March 2011

Contents

Preface 5

Introduction and summary 7

Chapter 1

The Price of Prejudice 13
Morten Hedegaard and Jean-Robert Tyran

Chapter 2

Correlates and Consequences of Distributional Preferences: an Internet Experiment 67
Morten Hedegaard

Chapter 3

To See is to Believe: Common Expectations in Experimental Asset Markets 123
Stephen L. Cheung, Morten Hedegaard and Stefan Palan

Preface

This PhD thesis was written during my time as a PhD student at the Department of Economics, University of Copenhagen. I am grateful to the department in general and to the Centre for Experimental Economics (CEE) in particular for providing me with an inspirational environment and generous financial support. I owe my greatest thanks to my supervisor and co-author, Jean-Robert Tyran. His drive, intriguing ideas and encouragement have been an inexhaustible source of inspiration and motivation. He was never too busy to discuss new ideas, technical design details or latest results and we have spent countless hours engaged in academic discussions in coffee houses in Copenhagen and Vienna. I have benefitted tremendously from our cooperation and I am deeply grateful for having had the opportunity to work closely with him.

I also thank my other co-authors Stefan Palan and Stephen L. Cheung who visited me several times in Copenhagen while working on our joint project. I thank Mathilde Almlund, Jonathan Schulz and the CEE members for providing insightful comments, valuable discussions and pleasant company. During my PhD, I visited the University of Tilburg, the University of Innsbruck, Karl-Franzens University Graz and the University of Vienna. I would like to thank Charles Noussair, Arno Parolini, Stefan Palan and everyone at the Department of Economics, University of Vienna for their kind hospitality.

There is a monetary cost associated with conducting research in experimental economics. I am grateful to the Department of Economics, the Rockwool Foundation Research Unit and to the Carlsberg Foundation for very generous financial support. Without their support, the research presented in this thesis would not have been possible.

I have befriended fellow PhD students at the University of Copenhagen and around the world. I am grateful to all of them for making this process even more enjoyable than it would have otherwise been. Finally, I thank my family and friends for their support, encouragement and for bearing over with me when things got a bit nerdy.

Morten Hedegaard
Copenhagen, March 2011

Introduction and Summary

The three papers in this thesis are experimental and make contributions to three distinct topics of behavioral economics. As such, the papers can be read independently of each other. The first is about discrimination in the work place. The behavioral theory of "taste"-based discrimination is due to Becker (1957) who modifies the utility function by allowing agents to derive utility not just from their material payoff but also from the (e.g. ethnic) type of workers they interact with. For instance, employers might have an animus against particular types of workers. If this is the case, they would be willing to give up money to employ their preferred type. The first paper of this thesis estimates the distribution of such animus for a sample in Denmark. We design a novel experiment which enables us to exogenously assign random costs to discrimination which allows us to estimates how behavior is affected by changes in the price of discrimination.

The second paper is about distributional preferences. Behavioral theories that incorporate distributional preferences modify utility functions such that agents care not just about their own material payoff but also for (some element of the distribution of) the material payoff of others. An early example is Fehr and Schmidt (1999) who model inequality aversion by letting agents' utility depend on both own payoff and on the difference between own and other agents' payoff. The second paper of this thesis uses a nonparametric approach and modifies Kerschbamer's (2010) XY test to elicit distributional preferences in Denmark. This is the first time that distributional preferences have been measured for a broad sample of the Danish population.

The third paper is about information in asset markets. Experimental methods have been used extensively to study mispricing – in particular price bubbles – in asset markets. This literature has been pioneered by Smith, Suchanek and Williams (1988) who find that price bubbles are common even in transparent markets where the fundamental value of assets is common knowledge. Standard economic theory would suggest (no) trade at fundamental in such an environment, given an assumption of common knowledge of rationality (i.e. that all market participants are rational and that all participants know this). The third paper tests the effect of common expectations (of rationality) on mispricing in such markets.

All three papers in the thesis use experimental methods to collect empirical evidence. The reason for using experiments is that they allow us to control the environment in which subjects make decisions and to impose exogenous variations in these environments. Hence, by using experimental techniques we can shed light on questions that cannot be easily answered using naturally occurring data. Thus, economic experiments should be seen as a supplement to the standard econometric analyses. This is especially true in cases where answering a particular research question requires exogenous variation that does not occur naturally or if important information cannot be observed and, thus, cannot be controlled for. As the thesis demonstrates – and as the title reflects – the control of information is crucial in order to identify preferences. The reason is that behavior observed in the field often can be explained with a variety of different arguments. By using experimental methods, we control the information available to agents which allows us to rule out confounding explanations. Using a revealed preference approach and combining our experimental data with micro-econometric methods and simulations allows us to identify the preference of interest.

While the three papers in this thesis all present experimental evidence, the experimental methods used vary across the papers. In particular, the first paper reports evidence from a natural field experiment, the second paper from an artefactual field experiment carried out over the Internet and the third paper from a traditional laboratory experiment[1]. Different types of experiments have different pros and cons and all serve their own purposes. We use a natural field experiment (where subjects do not know that they participate in an experiment) to study discrimination as it is essential that participants are unaware that we observe them in order to reduce 'experimenter demand' effects arising from a desire to act in a politically correct way. We use an artefactual field experiment to collect evidence over the Internet in order to have participants from all walks of life participating from the comfort of their homes. Finally, we study the effect of information in experimental asset markets using a standard laboratory experiment with the normal student population (similar to the rest of this literature). The three papers are briefly summarized below and the interested reader is referred to the chapters themselves for the full details.

Chapter 1, "*The Price of Prejudice*" (joint with Jean-Robert Tyran) examines ethnic discrimination in the work place. We set up a natural field experiment to distinguish between statistical and "taste"-based ethnic discrimination in a labor market setting. Statistical

[1] According to the taxonomy of Harrison and List (2004).

discrimination can occur when agents have imperfect information and base their decisions on beliefs about group characteristics (e.g. average productivity). Rational beliefs may result in what we call accurate statistical discrimination. However, if beliefs are biased then decision makers unknowingly pay a price for having prejudiced beliefs. We refer to this situation as belief-driven prejudice. "Taste"-based discrimination refers to the case where agents have an animus against a particular type and knowingly pay a price in order to work with the preferred, but less productive, type. In this case, we say that the price of prejudice is animus-driven. The difficulty with identifying the different forms of discrimination is that researchers need to know not only the true productivity of job applicants but also the beliefs of the decision makers. This is rarely the case with naturally occurring data. To solve this problem, we set up an experiment where we can both measure productivity precisely and control the productivity information.

We find that "taste"-based discrimination is common (discriminators are on average willing to forego 8 percent of earnings to work with a person of their own ethnic type) but remarkably responsive to the price of prejudice, i.e. to the opportunity cost of choosing a less productive worker on ethnic grounds (our best estimate of an elasticity is -0.9). In addition, we find that accurate statistical discrimination fails to explain observed choices, and that taking ethnic prejudice (both animus- and belief-driven) into account helps to predict the incidence of discrimination.

Chapter 2, "*Correlates and Consequences of Distributional Preferences: an Internet Experiment*" examines distributional preferences in a large sample of the Danish population. To do so, we set up a large-scale internet experiment that enables us to have participants that differ in their background characteristics. Using the internet as a platform serves two purposes. First, it enables us to investigate the correlation between participants' distributional preferences and their backgrounds as both are heterogeneous (in contrast, the standard student population is very homogeneous in terms of background characteristics). Second, it reduces some potential experimenter demand effects. For instance, Levitt and List (2007) hypothesize that subjects act in a more pro-social way when the experimenter is present and physically observes their behavior. The anonymity of the Internet is thought to reduce such effects. We use Kerschbamer's (2010) XY test which is a test focused purely on outcomes and not intentions. Again, control of information is crucial as decision makers need to be aware that the other participants' behavior (or their intentions) can not affect outcomes. Finally, we

investigate the effect of distributional preferences on cooperative behavior as measured in a standard public good game.

We find that 89 percent of subjects behave in a way that is consistent with either efficiency maximization (32 percent), inequality aversion (23 percent), selfishness (20 percent) or maximin preferences (14 percent). Thus, while the XY test is very comprehensive and allows for the full set of nine different preference types, we find that only four types have strong empirical relevance. In addition, we find that gender, age, attitudes towards competition and fairness and test scores for IQ and cognitive reflection correlate with distributional preferences. We find that agents with non-selfish preferences contribute more to the public good than those with selfish preferences and that the effects are substantial (contributions increase with between 6 and 11 percent), even after controlling for beliefs. Finally, we find that taking distributional preferences into account explains almost half of the gap between observed behavior and the behavior that is predicted by standard economic theory. These findings can be interpreted as a first validation of the XY test as it is able to predict behavior for the same subjects in a different task.

Chapter 3, "*To See is to Believe: Common Expectations in Experimental Asset Markets*" (joint with Stephen Cheung and Stefan Palan) investigates the effect of information about others' expectations in asset markets. We use the standard framework of Smith, Suchanek and Williams (1988) which is known to cause mispricing with inexperienced subjects. In a behavioral setting, this can be due to speculation even if all traders are rational as some might think that others are not rational. We compare two (main) treatments in which all traders have to correctly answer a battery of control questions. The extensive set of control questions are thought to ensure that all traders are rational on the individual level, i.e. they have full understanding of the environment. The fact that everyone answers control questions is common knowledge in one treatment but not in the other.

We find that mispricing is essentially eliminated when traders are aware that all have answered control questions. However, mispricing is substantial (and comparable in extent to the baseline without the extensive set of control questions) when control questions are not common knowledge. Thus, uncertainty about the expectations of others cause mispricing even in a setting where all traders are rational. This finding is in line with the original hypothesis by Smith, Suchanek and Williams (1988).

References

Becker, G.S. (1957): *The Economics of Discrimination*. University of Chicago Press.

Fehr, E. and Schmidt, K. (1999): A Theory of Fairness, Competition, and Cooperation. *Quarterly Journal of Economics* 114(3): 817-868.

Harrison, G. W. and List, J. (2004): Field Experiments. *Journal of Economic Literature* 42(4): 1009-1055.

Kerschbamer, R. (2010): The Geometry of Distributional Preferences and a Non-Parametric Identification Approach. *Unpublished Working Paper*.

Levitt, S. D. and List, J. A. (2007): What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World? *Journal of Economic Perspectives* 21(2): 153-174.

Smith, V. L., Suchanek, G. L. and Williams, A. W. (1988): Bubbles, Crashes, and Endogenous Expectations in Experimental Spot Asset Markets. *Econometrica* 56(5): 1119–51.

Chapter 1

The Price of Prejudice

Morten Hedegaard and Jean-Robert Tyran

The Price of Prejudice

Morten Hedegaard and Jean-Robert Tyran[†]

March 2011

Ethnic prejudice can result in "taste-based" or "statistical" discrimination in the workplace. We disentangle the two types of discrimination in a field experiment. We find that taste-based discrimination is common but remarkably responsive to the price of prejudice, i.e. to the opportunity cost of choosing a less productive worker on ethnic grounds. In addition, we find that accurate statistical discrimination fails to explain observed choices, and that taking ethnic prejudice into account helps to predict the incidence of discrimination.

Keywords: field experiment, discrimination, labor market.

JEL-codes: C93, J71

[†] Hedegaard: University of Copenhagen, Department of Economics, Øster Farimagsgade 5, building 26, DK-1353 Copenhagen K. Morten.Hedegaard@econ.ku.dk.
Tyran: University of Vienna, Department of Economics, Hohenstaufengasse 9, A-1010 Vienna, and University of Copenhagen, Department of Economics, Øster Farimagsgade 5, DK-1353 Copenhagen K, and CEPR (London). Jean-Robert.Tyran@econ.ku.dk.
We gratefully acknowledge generous financial support by the Rockwool Foundation's Research Unit. We are particularly grateful to Research Director Torben Tranæs for sparking our interest in this topic by drawing our attention to weaknesses of traditional experimental approaches to discrimination in the labor market. We are grateful for Torben's ongoing encouragement and support for the novel experimental procedures we developed and implemented. We also thank Dirk Engelmann, Simon Gächter and Ernesto Reuben for helpful comments and student assistants Nete Daly, Pelle Flachs and Anne Schjellerup Olsen for their practical support in conducting the experiment.

Introduction

Public debate is rightly concerned with ethnic discrimination in the work place because of its adverse consequences for the discriminated and for society at large. Ethnic discrimination is unfair to the discriminated, discourages investment in human capital and can lead to unemployment and even social unrest. Yet, this paper is not concerned with the adverse consequences of discrimination but with the economic causes of discrimination. In particular, we study the "price of prejudice" that discriminators pay for discrimination. Knowing whether discriminators pay such a price deliberately or unintentionally, and how they react to changes in that price is of utmost importance in designing effective policies to clamp down on discrimination.

Most of the economic literature has focused on "statistical discrimination" (Phelps 1972, Arrow 1973). This type of discrimination occurs if employers have imperfect information about the individual productivity of job candidates but can observe a group characteristic like ethnicity. If average productivity is indeed different across ethnic groups, an employer maximizes average profits at given wages by choosing the worker of the more productive ethnic group even if the individual job candidates are otherwise identical on observables. The literature usually assumes that employers form accurate judgments about the relative average productivity of workers by ethnic groups. Whether or not one likes to call such "accurate statistical discrimination" (ASD) prejudiced[1], it is clear that employers on average do not pay a price of prejudice when engaging in ASD.

In this paper, we investigate "prejudice" with two distinct meanings. First, ethnic prejudice can be belief-driven and result in inaccurate statistical discrimination. The employer may unintentionally pay a price of prejudice if he has false beliefs about average group productivities or about the ability to collaborate on the job when workers have different

---

[1] Economist would typically not call accurate statistical discrimination prejudiced while laypeople (and legislators) often do. In the theory of statistical discrimination majority employers are not assumed to have any ethnic animus against minorities nor are they assumed to have biased beliefs about the average performance of minorities on the job. In that sense, it seems appropriate to say that employers are not prejudiced. However, employers do choose between individuals based on the (true, relative) productivity of ethnic groups. If the distributions of productivities for minority and majority workers are different but overlap, it happens with some probability that the employer does not hire the most productive candidate. In that sense, the employer can be said to be prejudiced against (highly productive) minority individuals. Note that statistical discrimination in the work place is illegal in most countries.

ethnicity. Second, ethnic prejudice can be animus-driven resulting in "taste-based" discrimination (Becker 1957). For example, an employer may correctly believe that minority workers are on average more productive than majority workers but may dislike minority people for reasons that are unrelated to their productivity. Such an employer deliberately pays a price for his ethnic prejudice when hiring a majority worker at given wages.

The vast literature on ethnic discrimination in the work place (see Altonji and Blank 1999 for a survey) has struggled for decades with measuring the relevance of these two types of ethnic prejudice, essentially because beliefs and preferences cannot be directly observed. Experimental economists have developed tools to elicit beliefs and infer preferences from observed behavior in the laboratory but such measurement is fraught with difficulties if subjects are aware of being observed because of the illicit nature of discrimination. Researchers have developed clever designs (so-called correspondence tests and audit studies) to circumvent that problem using natural field experiments in which potential discriminators are not aware of being observed (e.g. Bertrand and Mullainathan 2004). However, such designs are not well suited to control the price of prejudice and, thus, to study how this price shapes discrimination choices.

We use a natural field experiment with two treatments to investigate the two types of prejudice and how they translate into hiring choices at given wages. In treatment Info, we identify taste-based discrimination by controlling for beliefs and by randomly assigning a price of discrimination. In essence, only animus-driven prejudice can matter for discrimination in Info because decision makers know the ethnicity and individual productivities of the job candidates. We randomly vary the price of discrimination by giving decision makers the choice between candidates of different productivities which allows us to estimate how taste-based discrimination responds to changes in its price. In treatment NoInfo, both taste-based and belief-based prejudice can matter for discrimination because decision makers do not know candidates' individual productivities and thus have to form beliefs about the average productivity of ethnic groups. In NoInfo, we investigate the relative predictive power of animus- and belief-driven prejudice by explaining the gap between observed behavior and the benchmark of "accurate statistical discrimination". To account for this gap, we elicit beliefs and use our estimate of taste-based discrimination from treatment Info. In both treatments, we take great care to run a natural field experiment, i.e. an experiment in which participants are not aware that they are in an experiment, to avoid possible bias in the measurement of morally sensitive (and illegal) ethnic discrimination.

The experiment proceeds as follows. We hire 169 juveniles from secondary schools in Copenhagen, Denmark, with Danish-sounding and Muslim-sounding names to pack letters for a large mailing and pay them at a piece rate. Workers are requested to show up for work twice in two consecutive weeks. In the first round, they work in isolation and we measure their individual productivity on the job. Before they come back for the second round, we call randomly selected workers on the phone and inform them that they will again do the same job but now have to work in teams of two. They are informed that they are paid according to the same piece rate as in round one and share earnings from team output in round two with the other team member. These randomly selected workers can choose whom to work with. The choice is between a candidate from the ethnic majority group and a candidate from an ethnic minority group. In treatment Info, we provide the decision maker with information about the individual productivity of the two candidates, i.e. the number of letters packed in round one, and their first names as a marker of ethnicity. The candidates are randomly selected from the pool of workers and therefore have random productivity differences. Treatment NoInfo is the same as Info except that decision makers are not informed about candidates' individual productivities. We elicit beliefs about individual and team productivity on a different but similar sample. We use these beliefs to test if beliefs are accurate and to evaluate how much of the price of prejudice can be attributed to biased beliefs and animus, respectively.

In treatment Info, we find that taste-based discrimination is common even at a substantial cost and that the tendency to discriminate is not different across ethnic types. We find that discriminators are on average willing to forego 8 percent of their earnings in round two to work with a person of their own ethnic type. Our main result from treatment Info is that taste-based discrimination is surprisingly responsive to the price of prejudice. Our best estimate is an elasticity of -.9, i.e. we find that the probability to discriminate falls by about 9 percent if the price of discrimination goes up by 10 percent. These results suggest that policies aiming to clamp down on animus-driven discrimination by increasing the price of discrimination to employers may be rather effective.

In treatment NoInfo, we find that accurate statistical discrimination (ASD) fails to explain observed outcomes since we observe a large gap between observed earnings and earnings with ASD (about 4 percent of total output). To account for animus-driven prejudice, we use our estimate from treatment Info. To account for belief-driven prejudice, we use elicited beliefs. At least 40 percent of that gap is explained by animus-driven prejudice alone, and at most one third is explained by belief-driven prejudice alone. Jointly, the two types of

prejudice explain about 60 percent of the gap. Thus, our results suggest that belief-driven and animus-driven ethnic prejudice are important causes of ethnic discrimination in the workplace, and need to be taken into account above and beyond the theory of accurate statistical discrimination.

The paper is organized as follows. Section 2 discusses related literature, section 3 describes our experimental design, and section 4 presents the results. Section 5 summarizes and concludes.

2 Measuring ethnic discrimination in the work place

The traditional econometric approach to measuring the effects of discrimination is to estimate a "wage gap" between a minority and a majority group based on observables such as education or years of experience on the job (see Altonji and Blank 1999 for a survey). However, attributing the entire unexplained part of such regressions to discrimination is problematic, mainly because the true economic value of a worker (the marginal product of labor) is not observed by the researcher. Such approaches only allow for indirect inference of whether discrimination is taste-based or driven by false beliefs, and such inference is fraught with difficulties. For example, List (2006: 19) notes that "An important lesson learned from the vast literature on discrimination is that data availability places severe constraints on efforts to understand the nature of discrimination, forcing researchers to speculate about the source of the observed discrimination."

Field experiments[2] circumvent this difficulty and have been used for more than 40 years to investigate the causes of ethnic discrimination in the work place (Daniel 1968 and Jowell and Prescott-Clarke 1970 are early examples. See Riach and Rich 2002 for a survey). Such field experiments traditionally come in one of two guises. First, in in-person audit studies, "testers" (i.e. actors) of different ethnicity are matched into pairs with respect to physical appearance and are trained to behave similarly in job interviews. For example, Pager, Western and Bonikowski (2009) study hiring in the low-wage labor market in New York and find that

[2] There is also a considerable literature on discrimination using laboratory research both in psychology and economics (e.g. Gneezy and Fershtman 2001 or Holm 2001; see Anderson, Fryer and Holt 2006). Field experiments have also been used to measure gender discrimination (e.g. Goldin and Rouse 2000) and other types of discrimination in the labor market (e.g. Neumark et al. 1996), and discrimination in other markets (e.g. Ayres and Kenny 1995, Levitt 2004, List 2004, Yinger 1998).

black applicants are about half as likely to receive callbacks or job offers as white applicants. Second, in correspondence tests, pairs of fictitious resumes are submitted to employers by mail. Discrimination is inferred from differential callback or job-offer rates across pairs of workers which are similar in all respects except for ethnicity. These approaches have the advantage of using controlled variation to isolate the causal effect of ethnicity on employers' responses (see List 2006 for a discussion). Controlling for productivity differences by making pairs of ethnically diverse candidates as similar as possible is appealing since observing unequal treatment of otherwise identical workers is closely tied to a common definition of discrimination.[3]

Despite their clear advantages, correspondence tests and in-person audits also have some limitations (see Pager 2007 for a discussion). A concern with in-person audits is that testers are usually informed about the purpose of the study which may induce them, perhaps unconsciously, to behave in ways that can distort findings. Another concern with in-person audits is that testers may differ in characteristics that seem relevant for their labor productivity to the employer but are not observed by the researcher (e.g. Heckman 1998). Essentially, the problem is that ethnicity cannot be randomly assigned to testers. This problem is circumvented by correspondence tests which make (fictitious) applications similar in the eyes of employers. But this strength is also a weakness of this approach. Since applicants are equally productive by design, discriminators do not pay a price for their prejudice and correspondence tests may therefore exaggerate the true extent of discrimination (e.g. Heckman and Siegelman 1993). In addition, correspondence tests are silent on how discrimination responds to changes in the price of discrimination because they usually do not vary the cost of choosing one candidate over the other (see Neumark 2010 for a discussion).

Bertrand and Mullainathan (BM, 2004) is an excellent example of a correspondence test.[4] BM submit pairs of resumes to job openings in Chicago and Boston. The pairs of

---

[3] Altonji and Blank (1999: 3168) define discrimination in the labor market as "a situation in which persons who provide labor market services and who are equally productive in a physical or material sense are treated unequally in a way that is related to an observable characteristic such as race, ethnicity, or gender".

[4] Correspondence tests are available for about a dozen countries, and they yield, by and large, evidence of pronounced discrimination. For example, Carlsson and Rooth (2007) find callback rates are 50 percent higher for applicants with Swedish-sounding names compared to Middle Eastern-sounding names in Sweden, Oreopoulos (2009) finds that callback rates are 40 percent higher for applicants with English-sounding names than with Chinese-, Indian- or Pakistani-sounding names in Canada.

resumes are carefully matched such that they are as similar as possible with respect to productivity signals while keeping them distinct in a formal sense to avoid that employers realize that the resumes are fictitious. BM use a typical "White-sounding" and a "Black-sounding" name in each pair as a marker of ethnicity. BM find that applicants with White-sounding names are about 50 percent more likely to receive call-backs than applicants with Black-sounding names. In addition, of the 157 employers who responded asymmetrically to White-sounding and Black-sounding applications, 83 favored White-sounding applications while only 39 favored Black-sounding ones.

A particularly innovative aspect of BM is their ability to benchmark the returns of having a White-sounding name. BM submit four resumes to each job opening, two similar ones of low quality and two of high quality. Quality differs along ten dimensions, for example with respect to years of experience or computer skills. This variation in quality allows BM to estimate "the return to a White name" which is found to be "equivalent to about eight additional years of experience" (p. 998). While such equivalents can be interpreted in terms of cost of discrimination to the discriminated they are difficult to interpret in terms of the price of prejudice paid by the discriminator which is the focus of our study.[5] The reason is that the opportunity cost of hiring a White worker of lower quality rather than a Black worker of high quality is not known to the researchers. While BM's main finding of a racial gap in callbacks is consistent with both animus-driven and belief-driven prejudice, the authors argue that neither theory can satisfactorily explain the full set of findings.

Both in-person audits and correspondence tests have important advantages, but concentrate on measuring the extent of discrimination when discrimination is free for the discriminator. Instead, we observe discrimination when there is a price to pay for being prejudiced, i.e. when discrimination is costly to the discriminator.[6] Our approach allows us to

---

[5] Caruso et al. (2009) use a related technique, so-called conjoint analysis, to estimate how decision makers trade-off relevant (like education and IQ) and irrelevant (body weight) characteristics in choosing a team-mates for a hypothetical trivia contest. While the paper estimates a trade-off, it is silent on the price of prejudice paid by the discriminator because the choices were not consequential.

[6] Few studies have been able to relate variations in price to discrimination choices in a context not related to work. For example, Baccara et al. (2009) use variation in the cost of adopting children in the US to estimate the willingness to pay for babies with particular (ethnic, among others) characteristics. Pope and Sydnor (2011) use variation in interest rates in online peer-to-peer lending to show that statistical discrimination of

put a price tag on discrimination choices or, borrowing Gary Becker's (1957) expression[7], to estimate how discrimination responds to the "price of prejudice", rather than just observing that discrimination occurs when it is costless.

3 Experimental design

A general description of the experiment is as follows. We recruit an approximately balanced sample of juveniles with Danish-sounding and Muslim-sounding names from secondary schools in central Copenhagen for a letter packing job. Volunteers commit to show up twice for packing letters and indicate their availability for work. In the first round, they pack letters at a piece rate in isolation. This round serves to measure individual productivity on the job. In the second round, workers are required to work in teams of two, and some randomly selected workers (the "decision makers") can choose their partner. We construct triples of workers by randomly drawing one decision maker and two "candidates", one with a Danish-sounding name and one with a Muslim-sounding name.

The discrimination choice is made between rounds one and two. We call the decision makers on the phone and explain that they will do the same job at the same piece rate in round two, but have to work in teams of two. In treatment Info, they learn the first names and the productivity (i.e. number of letters packed in round one) of the two candidates. In treatment NoInfo, they only learn the first names. In both conditions, decision makers know that all candidates are equally experienced and have similar characteristics. In particular, they know that all candidates have worked on the same job under identical conditions and that they are recruited from secondary schools. When the decision maker has made a choice, we call the chosen candidate requesting him or her to show up at a particular time. In round two, teams are formed according to the choices of the decision makers whenever possible, and workers are paid out for both rounds.

We took great care to implement a proper natural field experiment – in which participants are not aware that they are part of an experiment. In particular, we have been

---

black borrowers absent animus cannot explain net returns observed in loan-performance data. Levitt (2004) uses data from a TV show to test how statistical discrimination of candidates reacts to changes in cost.

[7] "Price and Prejudice" is the title of part 2 in Becker (1976, The economic approach to human behavior) which is a revised version of his PhD thesis, published in 1957.

careful at all stages of the experiment to assure that the job itself and the work conditions appear natural to participants, that the experiment (in particular the information provided to decision makers) is tightly controlled, and that all aspects of the experiment are consequential and do not involve deception.

Detailed description of procedures

Recruiting. We distributed hundreds of flyers in eleven upper secondary schools in central Copenhagen.[8] The flyer explains that the University of Copenhagen is looking for part-time workers to prepare a major mailing for research purposes. The flyer also explains that applicants are expected to show up for two hours in each of two consecutive weeks. Applicants are requested to call us on a phone number indicated on the flyer.

We recruited in upper secondary schools because these juveniles have relatively low outside options, are similar with respect to age (16-20 years old) and education, are legally allowed to work for money, and because there is considerable naturally occurring ethnic heterogeneity in this group (23% of juveniles in these schools are immigrants). Using a homogenous subject pool has the advantage of minimizing unobserved heterogeneity across ethnic types, for example with respect to language skills. In addition, it is feasible to recruit an approximately balanced sample by gender from this pool. The reason for wanting a balanced sample is that we keep the triples (see below) separate by gender to avoid confound of ethnicity and gender.

Names as markers of ethnicity. Upon calling us, we record the applicants' names, phone numbers, and where they saw the flyer. Applicants indicate when they are available for work in both rounds and are requested to make a commitment to show up at any of these slots. We classify the applicants according to their first names as Danish-sounding or Muslim-sounding. We call 169 persons with high availability[9], with names apt to evoke ethnic stereotypes, and in approximately balanced proportions (see table 1).[10]

---

[8] The flyer is reprinted in appendix A. Appendix B shows the location of the schools.

[9] 95 percent (n = 169) of participants were available 3 or more days, 55 percent on 6 or more days in round 2.

[10] Table 1 shows that the names of 7 workers did not fit either ethnic type. These workers (and the teams they worked in) are excluded from our analysis below. Table 1 does not list 27 workers who participated in a pre-test. These workers were recruited from a school where we did not recruit for the main experiment.

We called applicants with typical Danish-sounding and Muslim-sounding names because these ethnic groups are by far the largest in Denmark.[11] We use first names as markers of ethnicity since it is natural to refer to a person in Denmark by first name across all social strata. Using first names to evoke stereotypes is common practice in correspondence tests. These tests use fictitious first names which can be chosen to be particularly strong markers of ethnicity (e.g. Lakisha vs. Emily in Bertrand and Mullainathan 2004). In contrast, we use participants' actual first names to mark ethnicity. In a follow-up study with 144 subjects, we find that our ethnic markers are highly effective and confound rarely occurs. For example, names we classify as Muslim-sounding are thought to be Danish-sounding only in about 1 percent of the cases (see appendix D for details).

Note that the first names of the ethnic minority group are both Muslim-sounding but also foreign-sounding to Danish ears. Thus, our study cannot not provide a definitive answer on whether the animus we measure among Danes is directed at Muslims or foreigners living in Denmark more generally. However, a correspondence tests designed to investigate this issue (Adida et al. 2010) for France suggests that animus against Muslims is more pronounced than animus against foreigners in general.[12]

Table 1: Number of workers in round 1 by gender and ethnicity

	Ethnicity			
Gender	Danish-sounding name	Muslim-sounding name	Other name	*Total*
Female	40	46	5	*91*
Male	40	36	2	*78*
Total	*80*	*82*	*7*	*169*

[11] According to official statistics (2009, www.statistikbanken.dk), 69 percent of immigrants in Denmark are from non-Western countries, and most of these originate from countries with high proportions of Muslims such as Turkey, Iraq and Pakistan.

[12] The study combines a foreign-sounding last name (Diouf, a typical name in Senegal) either with a Christian (Marie) or Muslim (Khadija) first name. Response rates for Marie Diouf and a reference candidate with a typical French name (Aurélie Ménard) were not different. However, response rates for Khadija Diouf were significantly lower than for Marie Diouf.

Measuring individual productivity. A total of 169 persons work in round 1 of our experiment. Workers are requested to show up at particular times and are led to separate rooms to minimize interaction between them. The letter packing job is explained and demonstrated to each worker individually. The job involves packing letters marked with an ID-number. These numbers have to be looked up in a binder and are associated with different letter types. Depending on the type, letters have to be complemented with a gift and sorted into specific bins (see appendix C for details). When participants indicate that they understand the task, the payment scheme (the piece rate is DKK 4, approx. 0.5€ per letter), and that they are ready to start, an alarm clock is set in the control room (see Figure B2 in appendix B). After exactly 90 minutes a staff person returns to the worker and counts the number of letters packed. Each worker got a receipt confirming their entitlement and was paid at the end of round 2 to provide them with incentives to return.

The letter packing job is ideal for our purposes for several reasons. First, the job is easy to explain and easy to learn for juveniles within the given time frame. Second, the job can meaningfully be done both in isolation and in a team of two workers. Third, teamwork on the job requires minimal spoken interaction which minimizes the motive to discriminate against members from a different "speech community" (e.g. Lang 1986). Fourth, the task produces sufficient variation in individual output which is essential to make discrimination costly. Fifth, the job is not artificial. It is not unusual for juveniles to work in a temporary job like letter packing and the job is real in the sense that we effectively used the letters packed for a large-scale mailing.[13]

Matching procedure. Upon completion of round 1, we match workers into triples as follows. We randomly select a person to be the decision maker. Thus, the decision maker may have a Danish-sounding or a Muslim-sounding name. We then determine the set of all suitable candidates for this decision maker. This is the set of participants who are of the same gender as the decision maker, are from a different school, and are available for work on at least one of the time slots indicated by the decision maker. We randomly draw two candidates from this set. One candidate is of the same ethnic type as the decision maker (*same* for short),

---

[13] We used the letters packed for a mailing to recruit participants for a large-scale internet study. This study used different letter types necessitating sorting the letters. We randomly checked 5 letters for each participant in round 1. The error rate was low (0.05) and did not differ by ethnic type ($p = 0.270$, χ^2-test). Error rates also do not differ by team composition in round 2 ($p = 0.688$, χ^2-test).

one is of the other type (*other* for short). In treatment Info, the draws are repeated until *same* is *less* productive than *other* and the two candidates are available on different weekdays. If no such pair exists, we randomly draw a new decision maker from the pool.

We randomly draw decision makers to observe discrimination choices by both ethnic types. The ability to observe discrimination choices by minority decision makers is, to the best of our knowledge, a unique feature of this study. For example, correspondence tests usually do not observe the ethnicity of the employer and simply assume that he or she belongs to the ethnic majority. We match candidates and decision makers from different schools to exclude that they personally know each other, thus avoiding confound of ethnic discrimination with a preference for a personal acquaintance. We are able to match teams from different schools because we gather information about school affiliation from participants when they apply for the job over the phone. Randomly drawing two candidates serves to generate a random price of discrimination (i.e. the earnings foregone by choosing *same* over *other*). Randomly matching the candidates with a decision makers serves to make price independent of any animus that may be present. Random assignment of price to decision makers is a precondition for identifying taste-based discrimination, as is discussed in more detail in section 4.1 and appendix G. The restriction imposed in Info that *same* has lower productivity than *other* serves to maximize the number of informative choices. Choices are informative in the sense that decision makers with strong animus are likely to be detected. The reason why the candidates must be available on different weekdays is that we frame the discrimination choice as a choice between two weekdays rather than between two persons, as is explained next.

Discrimination choice. The discrimination choice is made on the phone prior to round 2. Upon answering the phone, the decision maker is asked to confirm availability on the two time slots determined by the matching procedure (Tuesday and Wednesday 2 p.m. - 4 p.m., say). If the decision maker cannot reconfirm availability, we say we have to make new arrangements and call back later. In this case, the triple pertaining to this decision maker is reinserted into the pool and a new triple is drawn according to the matching procedure described above. If the answer is affirmative, decision makers are informed that the job in round 2 is the same and is paid according to the same piece rate as in round 1. They are told that, unlike in round 1, they have to work in teams of two and that they have to share the

revenue from teamwork.[14] Decision makers are told that which person they are going to work with depends on which day they choose. In treatment Info, the decision makers are told the first names and the number of letters packed in round 1 for both candidates and asked to make a choice. For example, "If you choose Tuesday, you will work with Ahmed who packed 150 letters last week. If you choose Wednesday, you will work with Christian who packed 110 letters last week. So, when would you like to work, Tuesday or Wednesday at 2 p.m.?" In treatment NoInfo, the procedure is the same except that we do not mention the individual output of candidates in round 1.[15]

An important advantage of this procedure is the high degree of control it provides over the information available to the decision makers. In both treatments, decision makers know that candidates are similar (they are recruited from the same set of schools) and have the same experience on the job (they all worked in round 1 under the exact same conditions). Beyond that, in treatment Info, the decision makers know *only* the names and productivities of the candidates. Since they cannot personally identify or see the candidates, factors such as attractiveness or personal appearance cannot affect decisions in our design (see e.g. Möbius and Rosenblat 2006 for experimental and Hamermesh and Biddle 1994 for field evidence on personal attractiveness and discrimination). We frame the discrimination choice as a choice of workdays rather than persons to minimize so-called Hawthorne or experimenter demand effects (see Zizzo 2010 for a general discussion). Such effects may result from participants' concerns to conform with notions of political correctness (see e.g. Kawakami et al. 2009).

Credibility and consequentiality. We take great care to create a natural setting, to measure output and to provide information with tight control, to insure that all information provided to decision makers is truthful, and that choices are consequential. For example, decision makers were presented with a choice between two real people, we indicate their actual first names, and their actual productivity in round 1. Decision makers are matched to work with the partner of their choice in round 2 whenever possible (i.e. when both show up on time) which implies that the chosen candidate cannot make a discrimination choice.

---

[14] If asked, we justified that the job has to be done in teams of two by explaining that "we found out that working in teams of two is more effective and therefore workers on average earn more than last week". We knew from a pretest with 27 participants that this claim is true.

[15] Non-chosen candidates were reinserted into the pool of participants and were matched into another triple, either as decision maker or candidate. Thus, our design does not necessarily imply a cost to the discriminated.

We believe that our choice of the location and work task was highly credible in the sense that workers did not know that they were participating in an experiment. We made choices consequential for two reasons. First, consequentiality serves to avoid deception and disappointment. For example, decision makers who opt for a highly productive co-worker would be antagonized if forced to work with a low-productive partner in round 2. Second, the ability to observe team output in round 2 allows us to identify taste-based discrimination in treatment Info by controlling for a particular type of (team-work related) ethnic prejudice, as is explained next.

Team output in round 2 can in principle depend on the individual productivity of team members and the ethnic mix of the team. Because we observe output in round 2, we can test if that is the case. We show in section 4.1 that team output is very much driven by individual productivity in round 1, but does not depend on the ethnic team composition. That is, we find that the team production function is not "type specific". But decision makers may falsely believe that it is. For example, despite being told in Info that *same* is less productive when working in isolation than *other*, a decision maker may believe that team output is higher when working with *same* rather than *other* because "different types don't work together well". If so, observing a choice of *same* over *other* in Info is not an indication of taste-based discrimination but of a false belief. Therefore, we need to control for beliefs about the type-specificity of the team production function to correctly identify taste-based discrimination. Section 4.3 shows that the team production function is not believed to be type-specific. Since the team production function is in fact not type specific and is not believed to be, beliefs are correct in this dimension, and choosing *same* over *other* in Info is therefore a clear indication of taste-based discrimination.

4 Results

Section 4.1 estimates the price of taste-based discrimination, i.e. the earnings foregone by choosing *same* over *other* in Info, using a team production function. We find that the team production function is not type-specific which implies that discriminators pay a positive price for discrimination (i.e. choosing *same*) in Info. We argue that this price is known to decision makers since round 1 productivity is an excellent predictor of the price.

Section 4.2 presents the results for treatment Info. We find that 38 percent of decision makers engage in taste-based discrimination and they pay a price of about €5 on average.

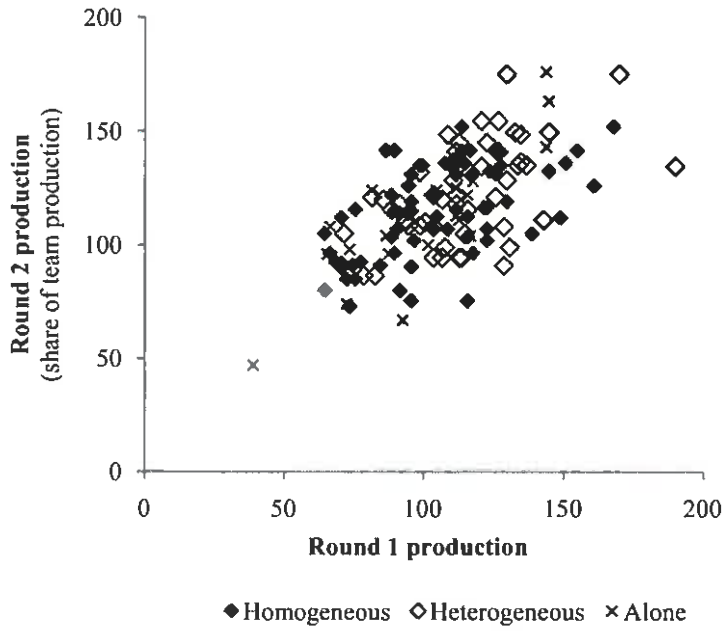
Importantly, we find that the probability to discriminate falls with its price and that the tendency to discriminate is not different across ethnic types after controlling for its price. We also estimate a distribution of the willingness to pay for taste-based discrimination.

Section 4.3 presents results for treatment NoInfo, i.e. when both animus-driven and belief-driven prejudice can matter for discrimination. We report results from a complementary study eliciting beliefs on production. We find that the team production function is not thought to be type specific, i.e. we find no evidence for prejudice about the ability to collaborate across types. However, we find that beliefs are biased in the sense that true productivity differences across ethnic types are underestimated. We find that accurate statistical discrimination (ASD) fails to account for observed choices. In fact, observed earnings are lower than those implied by ASD and about 60 percent of that gap is accounted for by animus and biased beliefs jointly. Thus, our estimate of taste-based discrimination from Info together with elicited beliefs predicts observed choices much more accurately than ASD.

4.1 The price of taste-based discrimination

We define the price of taste-based discrimination to the discriminator in Info as earnings foregone by choosing a less productive co-worker of the same ethnic type rather than a more productive worker of the other ethnic type. To measure this price, we estimate a team production function showing how workers with particular productivities in round 1 map into output of ethnically homogeneous and heterogeneous teams in round 2. We then estimate for each decision maker the marginal product of labor for the two candidates. This analysis yields the important result that team production is not type-specific, i.e. that two workers with given individual productivities produce the same output independent of the ethnic composition of the team. This result implies that a decision maker has a clear monetary incentive to choose the more productive candidate which, in treatment Info, is by design *other*. In other words, there is a price to pay for choosing *same*. If this price is known to the decision maker, a choice of *same* is a clear indication of taste-based discrimination, assuming utility maximizing choices. We argue that decision makers had almost perfect knowledge about the price in Info because candidates' round 1 productivities are excellent predictors of this price.

Figure 1: Production in round 1 and round 2



Note: The figure shows the number of letters packed in isolation in round 1 and the share of letters packed in round 2 for individuals who worked in round 2 in homogeneous teams (black diamonds, n = 68), in heterogeneous teams (white diamonds, n = 52), or alone (crosses, n = 20). The share is 50% of team output for those working in teams, and 100% of individual output for those working alone.

Figure 1 shows a scatterplot of worker i's share of production in round 2 (i.e. half of the letters jointly packed) by production in round 1 (i.e. letters packed in isolation). Black diamonds represent individuals in heterogeneous teams (52 individuals) and white diamonds represent individuals in homogenous teams (36 both Danish-sounding, 32 both Muslim-sounding). Crosses represent individuals working alone in round 2 (20 individuals) because they or their partner did not show up on time. The figure shows that there is considerable variation in both round 1 production (the average is 107 letters packed, sdv = 24) and in round 2 (average 115, sdv = 24). As expected by virtue of random treatment allocation, decision makers' distributions of round 1 production are not different across treatments (p = 0.528, Kolmogorov-Smironov test). Workers with Danish-sounding names tended to be more productive in round 1 than those with Muslim-sounding names (116 vs. 100, p = 0.000, Mann-Whitney test). This finding has important implications for our analysis in both treatments and is discussed in detail below.

We estimate the team production function using all observations of workers who completed both rounds[16] as

$$\ln\left(Y_{i,j}^{i}\right) = \beta_0 + \beta_1 \cdot \ln x_i + \beta_2 \cdot \ln x_{j,j\neq i} + \beta_3 \cdot \ln x_i \cdot Alone + \gamma \cdot \mathbf{X} + \varepsilon_i ,$$

where $Y_{i,j}^{i}$ is worker i's share of the team output in round 2 when working with co-worker j. We estimate team production as a function of worker i's own production in round 1, x_i, the production of the co-worker j in round 1 (x_j), an interaction term to capture different learning effects when working alone (*Alone* = 1 and $x_j = 0$ if i is working alone in phase 2), and a vector of variables characterizing the team composition (e.g. by ethnic type).

Table 2 shows various estimates for the team production function. The positive coefficients in the first two lines show that teams tend to be more productive if their members have high productivity in round 1. The coefficients in the third line reflect learning by those working alone in round 2. These coefficients are very similar in size to the previous ones suggesting that there is no gain from specialization in our task since those who happened to work alone are on average equally productive as those working in teams.[17] The significant coefficient for *Male* shows that males are about 6 percent more productive than females in round 2. Taken together, round 1 output explains a considerable share of variation in team output (adjusted R^2 is about .61 in all specifications) which implies that the information available to decision makers is an excellent predictor for the price of discrimination.

---

[16] In total, 140 workers completed both rounds according to the description in section 3. Observations from teams with workers having names which do not fit either ethnic type are excluded from our regression.

[17] Average earnings are the same whether working alone or in a team in round 2, holding everything else constant ($p = 0.573$, t-test). The coefficients in the first three lines of specification A are very similar because the share of team output for worker i and j is the same (one half) by definition and the share for a worker i working alone is estimated assuming a team mate j with the same round 1 production as worker i.

Table 2: Team production function

Dependent variable: $\ln(prod_{2i})$	(A)	(B)	(C)	(D)
$\ln(prod_{1i})$	0.416***	0.408***	0.421***	0.419***
	(0.044)	(0.044)	(0.050)	(0.051)
$\ln(prod_{1j})$	0.416***	0.426***	0.421***	0.428***
	(0.044)	(0.045)	(0.050)	(0.050)
$\ln(prod_{1i}) \cdot$ Alone	0.416***	0.424***	0.324***	0.327***
	(0.044)	(0.044)	(0.107)	(0.109)
Male	0.064***	0.063***	0.064***	0.064***
	(0.022)	(0.022)	(0.023)	(0.023)
Decision maker		-0.018		-0.017
		(0.024)		(0.030)
Alone			0.452	0.468
			(0.545)	(0.549)
Danish-sounding team			0.037	0.041
			(0.025)	(0.033)
Muslim-sounding team			-0.019	-0.010
			(0.035)	(0.039)
Decision maker · Heterogeneous				0.012
				(0.045)
Constant	0.841***	0.843***	0.785**	0.768**
	(0.219)	(0.220)	(0.315)	(0.317)
Adj. R^2	0.611	0.610	0.615	0.610
N	140	140	140	140

Notes: Dependent variable is (the logarithm of) the number of letters packed in round 2 by worker *i* if working alone ($n = 20$) or, if working in a team ($n = 120$), half of the number of letters packed by *i*'s team. $prod_{1i}$ is the number of letters packed in round 1 by worker *i*, $prod_{1j}$ the number of letters packed by *i*'s co-worker in round 2. *Alone* is a dummy set to 1 if worker *i* works alone in round 2, *Male* is worker *i's* gender, *Decision maker* indicates if worker *i* makes a choice of co-worker. The remaining dummies characterize team composition in round 2. Numbers in parentheses are robust standard errors. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Model B adds the dummy variable *Decision maker*. The insignificant estimate suggests that selection is not a serious issue with respect to team production as decision makers (after controlling for individual productivities) do not have significantly different productivity from those who have no choice to make. This is true for decision makers in general as well as for decision makers selecting into heterogeneous teams (see interaction term *Decision maker · Heterogeneous* in model D). Models C and D add dummies for team composition to test if ethnically homogenous teams are more productive than heterogeneous teams (which is the

reference category in the regression). The insignificant estimates show that the team production function is not type-specific. That is, given individual productivities, heterogeneous teams are equally productive as homogenous teams.

Taken together, the estimates on the production function show that much of the variation in team production is explained by one's own productivity and the productivity of the co-worker (which are both known when making the choice), but essentially nothing is explained by the ethnic type of the co-worker. This finding is important because it implies a monetary incentive to choose *other* in treatment Info. In other words, there is a price a to pay for discrimination, and decision makers had all the required information to know the price.

The price of taste-based discrimination is defined as earnings foregone by choosing to work with *same* rather than *other* in treatment Info. This price is not directly observed in our experiment because the decision maker only works with the chosen candidate but not with the non-chosen candidate. We thus have to estimate the counterfactual. In estimating the price of discrimination, we use specification A in table 2 because all variables included in the other models are insignificant. The price is then the difference between decision maker i's earnings with *other* minus the earnings with *same*[18]

$$Price_i = p(\hat{Y}^i_{i,other} - \hat{Y}^i_{i,same}) > 0 \ \forall \ i \ .$$

We find that the distribution of $Price_i$ (mirrored on 0) is normal (p = 0.818, Shapiro-Wilk; p = 0.721, Shapiro-Francia; p = 0.901, Skewness/Kurtosis test for normality), as is expected by virtue of random sampling of candidates.

4.2 Taste-based discrimination

Section A) below shows that the probability to discriminate falls as its price increases. Section B) estimates the willingness to pay for taste-based discrimination.

Before proceeding to estimation, we provide some descriptive statistics. Decision makers in treatment Info all face a positive price of discrimination by design, on average €6.7 (sd = €4.7). We observe that 38 percent of decision makers in treatment Info choose to discriminate,

[18] The price of discrimination expressed in Euros is obtained by multiplying the difference in output with p which is the product of the piece rate (DKK 4 per letter) and the exchange rate (0.13 Euro per DKK).

i.e. choose *same*. This result is novel since we show that taste-based discrimination is common even when decision makers face a positive and known price of discrimination.

A first finding supporting our claim that higher (randomly assigned) prices causally reduce discrimination is that discriminators face lower prices on average than non-discriminators (€4.9 vs. €7.8). Both a Kolmogorov-Smirnov test ($p = 0.091$) and Wilcoxon rank-sum test ($p = 0.052$) show that prices are different for the two groups (see appendix G for tests showing that prices are randomly assigned). The average expected price of €4.9 for discriminators may seem low in absolute terms but is strikingly high in relative terms. For example, the average discriminator gives up 8 percent of round 2 earnings to work with *same* for 90 minutes.

A) The demand for taste-based discrimination

We estimate the demand for discrimination using a revealed preference approach. Assuming $Price_i$ is known to decision makers[19] and choices are utility maximizing, decision maker i reveals to have a "taste" for discrimination $a_i \geq Price_i$ if he chooses *same*. In this case, we say the decision maker engages in taste-based discrimination (and we assign a value $Discr = 1$). Conversely, the decision maker reveals to have $a_i < Price_i$ if he chooses *other*, and we say the decision maker does not discriminate ($Discr = 0$). Given a distribution of animus a in the sample, utility maximization implies that fewer decision makers prefer to discriminate as its price increases. In other words, the demand for discrimination is downward-sloping.

We regress the probability of observing discrimination ($Discr = 1$) on the price of discrimination as defined in section 4.1 (plus other controls explained below) as follows[20]

$$\Pr(Discr_i = 1 \mid \mathbf{X}) = \Phi(\mathbf{X}'\beta + \varepsilon_i) .$$

Model (1) in table 3 provides the most parsimonious specification showing that the law of demand holds for taste-based discrimination. The coefficient on *Price* shows that discrimination falls by 3.6 percent if the price of discrimination goes up by €1. Note that this number is our best estimate for the average marginal change. Due to the non-linearity of the

---

[19] Below, we use the estimation results from the team production function to calculate *Price*. This implicitly assumes that decision makers know the team production function. Appendix D shows that our results are robust to this assumption. In particular, Appendix D shows that using raw productivity differences between candidates in round 1 as a proxy for *Price* yields the same qualitative results as those reported in table 3.

[20] We report probit estimates throughout the paper. Logit regressions yield qualitatively similar results.

demand relation, this marginal effect is not informative for larger changes in cost. We provide estimates for such changes in the discussion of figure 2.

Table 3: The demand for taste-based discrimination

Dependent variable: Discr	(1)	(2)	(3)	(4)
Price	-0.036**	-0.035**	-0.034**	-0.038*
	(0.016)	(0.017)	(0.016)	(0.020)
Danish-sounding		0.020		-0.045
		(0.160)		(0.286)
Male		-0.056		-0.022
		(0.152)		(0.284)
Danish-sounding · Price			0.005	0.011
			(0.022)	(0.040)
Male · Price			-0.007	-0.004
			(0.018)	(0.036)
R^2	0.082	0.085	0.086	0.087
N	37	37	37	37

Notes: The table shows average marginal effects estimated from Probit regressions. Numbers in parentheses are robust standard errors. *Discr* = 1 for a decision maker choosing *same* and 0 otherwise. *Male* and *Danish-sounding* are dummies characterizing the decision maker. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

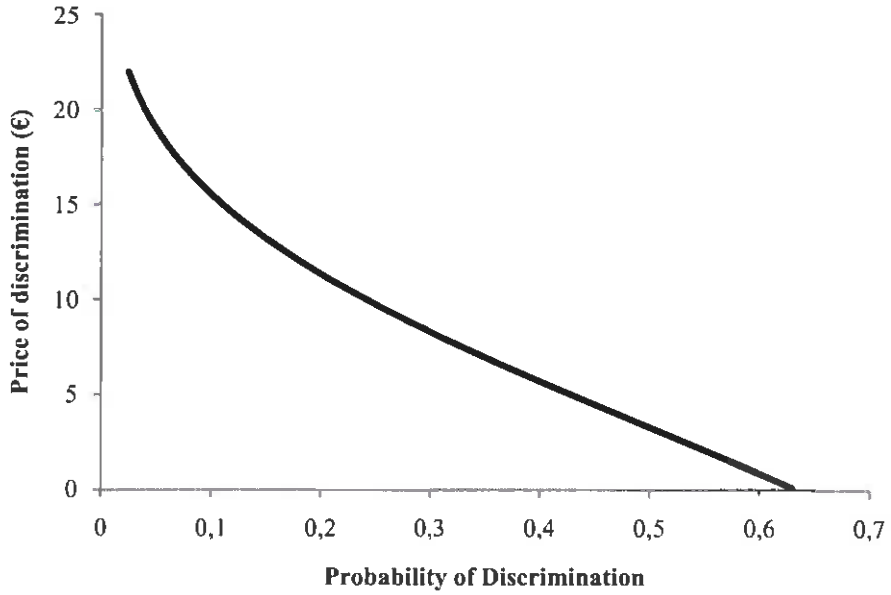
Model (2) adds dummy variables for gender (*Male*) and ethnic type (*Danish-sounding*) of the decision maker. The insignificant estimate on *Danish-sounding* indicates that the tendency to discriminate is not different across ethnic types, after controlling for differences in prices. We think that this is a remarkable result for two reasons. First, attention both in the literature and policy debates usually focuses on discrimination of the minority group by the majority group because the adverse consequences of discrimination (for the discriminated and society at large) are more pronounced in this case. In fact, members of the majority group are more often in the position to discriminate, and workers from the minority group tend to be disadvantaged. However, our results suggest that observing more frequent discrimination of minorities may be simply due to the fact that majority decision makers have more opportunities to discriminate rather than a stronger ethnic animus.

Second, this result highlights the importance of controlling for prices when measuring discrimination. From simply looking at discrimination percentages, a layperson may be misled to conclude that decision makers with Danish-sounding names are more likely to

discriminate. In fact, decision makers with Danish-sounding names discriminate in 44 percent of the cases, while those with Muslim-sounding names do so in only 33 percent of the cases (however, $p = 0.517$, χ^2 test). Yet, these differences do not reflect differences in animus because decision makers with Danish-sounding names face a lower price on average than decision makers with Muslim-sounding names (€5.2 vs. €7.8, $p = 0.078$, KS). The reason is that workers with Danish-sounding names are systematically more productive (116 letters) in round 1 than participants with Muslim-sounding names (100 letters). According to regressions (2) and (4) in table 3, these price differences explain the observed differences in taste-based discrimination across ethnic types (*Danish-sounding* is insignificant, but *Price* is significant).

Model (3) adds the interaction terms *Danish-sounding* · *Price*, and *Male* · *Price*. The respective estimates are insignificant, suggesting that responses to changes in price are not different across ethnic types and gender. Model (4) combines (2) and (3) and yields the same results.

Figure 2: The demand for discrimination



Notes: The figure shows the relation between the probability of discrimination (choosing *same*) in Info and the price of discrimination, calculated using specification (1) in table 3.

Figure 2 summarizes our main finding. Decision makers respond strongly to changes in prices. For example, the figure shows that increasing the price of discrimination by one standard deviation from the average (i.e. from €6.7 to €11.4) reduces the probability of discri-

mination by 45 percent (from .36 to .20). Conversely, decreasing the price by one standard deviation from the average (i.e. from €6.7 to €2.0) increases the probability by 54 percent (from .36 to .55). Another way to describe the remarkable price-responsiveness is to estimate an elasticity which indicates the percentage decrease in the probability to discriminate in response to a 1% increase in price. Our best estimate is -0.9. This elasticity is an average of all elasticities, evaluated at each observation. In conclusion, we find that the demand for taste-based discrimination is downward-sloping and is surprisingly elastic.[21]

B) Willingness to pay for taste-based discrimination

An alternative representation of our main finding is in terms of the willingness to pay for taste-based discrimination. According to the revealed preference approach described above, decision maker *i* reveals to have willingness to pay $a_i \geq Price_i$ if he chooses *same*. Conversely, the decision maker reveals to have $a_i < Price_i$ if he chooses *other*. We assume that willingness to pay is normally distributed in the population, $a_i \sim N(\mu_a, \sigma_a^2)$. We estimate μ_a and σ_a from estimated $Price_i$ (using model A in table 2) and observed discrimination choices as follows. We define the probability of discrimination as

$$\begin{aligned}\Pr(Discr = 1 \mid Price_i) &= \Pr(a_i \geq Price_i) \\ &= 1 - \Pr(a < Price_i) \\ &= 1 - F_a(Price_i)\end{aligned}$$

where F_a is the CDF of a. We use probit estimation to estimate this probability (see model 1 in table 3):

$$\Pr(Discr = 1 \mid Price_i) = \Phi(\beta_0 + \beta_1 \cdot Price_i + \varepsilon_i)$$

and use the estimates $(\hat{\beta}_0, \hat{\beta}_1)$ to obtain the distribution of the willingness to pay:

$$F_a(x) = 1 - \Phi(\hat{\beta}_0 + \hat{\beta}_1 \cdot x), x \in \Re$$

We find that the average decision maker in our sample is willing to pay μ_a = €3.2 to work with *same* rather than *other* (σ_a = €9.6). Our estimation approach allows decision makers to have positive (a dislike of *other*) or negative (a preference for *other*) animus.

---

[21] Interestingly, our estimate at a price of zero is close to the estimates in correspondence tests. For example, we find that a decision maker with a majority name picks *same* with a probability of 63 percent at a zero price. Bertrand and Mullainathan (2004) find that workers with White names are about 50% more likely to be called back which, assuming that employers are White, translates into a 60 percent probability of choosing *same*.

Interestingly, our estimate suggests that while a majority (63 percent) dislikes working with *other*, a considerable share also prefers working with *other*.

4.3 Discrimination when both types of prejudice can matter

In treatment NoInfo, decision makers do not know candidates' individual productivity but do know their ethnic types. Thus, decision makers need to form beliefs about the relative productivity of workers across types to make optimal discrimination choices. Differences in beliefs about relative productivity are likely to be mainly driven by ethnicity in our design. The reason is that decision makers know that all candidates have very similar age and educational background (because they are recruited from the same set of schools) and have the exact same amount of experience with the work task. We thus control information and make the candidates similar – except for their ethnicity – in the eyes of the decision makers.

Accurate statistical discrimination (ASD) assumes that decision makers form rational (i.e. on average correct) beliefs and that decision makers have no animus.[22] That is, ASD assumes profit-maximizing choices. ASD predicts that all decision makers in NoInfo choose the candidate with a Danish-sounding name because these workers are on average more productive (116 vs. 100 letters packed in round 1). We find that ASD grossly mispredicts choices. In fact, about half of the choices are for the less profitable type, and decision makers with Muslim-sounding names are particularly prone to make such choices.

Treatment NoInfo serves to evaluate the predictive power of ASD against animus-driven and belief-driven prejudice in explaining observed outcomes. Such a comparative test is demanding because it requires that the researcher measures animus and rational beliefs as well as the actual beliefs. Our study is ideally suited to measure rational beliefs, i.e. the true average price of discrimination, because we precisely estimate individual marginal products of labor using the team production function. Our design is less suited to directly measure whether decision makers have biased beliefs about the average price of discrimination. The reason is that we make every conceivable effort to implement a natural field experiment to avoid distorted responses from moral bias. Thus, asking participants in NoInfo directly about

[22] Altonji and Pierret (2001: 316) explain that they "are using the term 'statistical discrimination' as synonymous with the term 'rational expectations' in the economics literature. We mean that in the absence of full information, firms distinguish between individuals with different characteristics based on statistical regularities. That is, firms form stereotypes that are rational given their information."

their expected price of choosing one candidate over the other is not an option.[23] Instead, we elicit beliefs about the average price of discrimination indirectly, from a sample of similar juveniles. Eliciting beliefs serves two purposes: to test whether the team production is thought to be type-specific (see section A below), and to assess the relative explanatory power of animus-driven and belief-driven prejudice in accounting for observed discrimination in NoInfo (see section B).

A) Eliciting beliefs

We recruit $n = 353$ participants with Danish-sounding and Muslim-sounding names from secondary schools on the outskirts of Copenhagen where we do not recruit for the experiment.[24] We carefully describe the work task to participants and elicit beliefs about the productivity of individuals and teams. In particular, each participant is presented with the names of 7 randomly selected workers and 6 randomly selected teams, all of the same gender as the participant. Participants are asked to guess how many letters each worker packed in round 1 and each team packed in round 2. To benchmark their expectations, we inform participants about the median production in round 1 and round 2 (see appendix H for details). Participants are rewarded for guessing correctly (using a quadratic scoring rule[25]).

Figure 3 shows participants' average beliefs about production in round 1 and 2 by ethnic type of the participant. The horizontal axis shows beliefs about productivity differences *same* minus *other*. The vertical axis shows beliefs about productivity differences between homogeneous teams of the same type as the participant and heterogeneous teams. Each dot represents one participant. The figure shows that the black dots tend to be located slightly to the right and above the zero lines, while white dots tend to be located to the left and below the zero lines. These tendencies reflect the fact that participants of both types correctly tend to think that workers with Danish-sounding names are individually more productive, and that, as

---

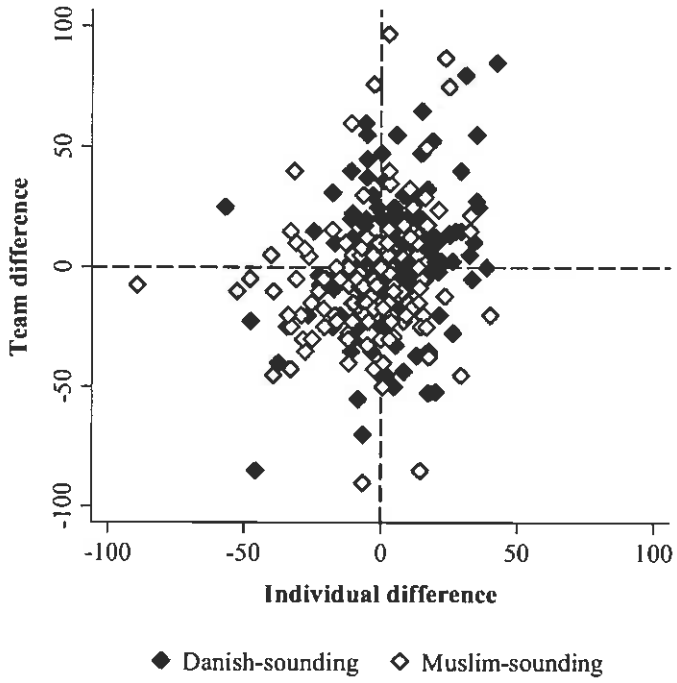
[23] People tend to be more prejudiced than they admit. For example, Kawakami et al. (2009: 277) show "that people's predictions regarding their emotional distress and behavior in response to a racial slur differ drastically from their actual reactions". Studies using survey-based data on animus may therefore yield lower-bound estimates for animus. For example, Charles and Guryan (2008) use 21 survey questions to argue that animus-based prejudice explains up to 25 percent of the racial wage gap in the US.

[24] We omit 42 persons who are classified as having "other" names from the analysis below.

[25] Participants receive $\max(0;\ 50 - 0.03d^2)$ where d is the difference between the true productivity and the guess. Average earnings in the belief elicitation study are €13.6.

a consequence, teams with two workers with Danish-sounding names are more productive than teams with two workers with Muslim-sounding names.

Figure 3: Beliefs about production in round 1 and 2



Notes: The horizontal axis shows the difference in average beliefs for numbers of letters packed individually in round 1 by workers of the same ethnic type minus the other ethnic type for participants with Danish-sounding (black dots) and Muslim-sounding names (white dots). The vertical axis shows the difference in average beliefs for number of letters packed in round 2 by homogeneous teams (both workers of the same ethnic type as participant) minus number of letters packed in round 2 by heterogeneous teams for participants with Danish-sounding (black dots) and Muslim sounding names (white dots). $N = 353$, of which 204 participants have Danish-sounding names and 149 have Muslim-sounding names. Two outliers with values > 100 are omitted from the figure.

Statistical testing confirms this visual impression. We find that participants have qualitatively correct beliefs in the sense that they believe that individual workers with Danish-sounding names pack more letters on average than workers with Muslim-sounding names ($p = 0.004$, Wilcoxon signed-rank test, WSR). However, average beliefs are quantitatively biased since the true difference across types of workers is larger than the expected difference (16 vs.

3 letters)[26]. In other words, participants underestimate the true productivity difference across types. Consistent with the belief that workers with Danish-sounding names are individually more productive, we find that teams with more Danish workers are believed to more productive.

Importantly, we find no evidence for ethnic prejudice in the sense that the team production function is *thought* to be type-specific. Our analysis in table 2 has shown that, after controlling for individual productivity, heterogeneous teams in fact are equally productive as homogeneous teams. The analysis below shows that participants do not think that workers earn more in a homogeneous team than a heterogeneous team, for given round 1 output. Put differently, neither do the juveniles believe nor do they have a reason to believe that selecting a co-worker of the same type is more profitable for given productivities of workers. To test, we regress

$$\Delta_i = \beta_0 + \beta_1 \delta_i + \beta_2 Danish + \varepsilon_i,$$

where Δ_i and δ_i capture the participants' beliefs about output of teams and individuals of different ethnic types. More specifically, Δ_i is participant i's belief about output in a homogenous team of the same type as i minus i's belief about output in a heterogeneous team. Thus, Δ_i captures how much participants with Danish-sounding names thought that all-Danish teams outperform heterogeneous teams, and vice versa for participants with Muslim-sounding names. The variable δ_i is i's belief about output of individual workers of the same type as i minus i's belief about output of workers of the other type. Thus, δ_i captures how much participants with Danish-sounding names thought that Danish workers outperform Muslim workers, and vice versa for participants with Muslim-sounding names. The dummy variable *Danish* equals 1 if the participant has a Danish-sounding name and is used to check whether the two groups differ in their beliefs about the production function.

The regression yields an insignificant coefficient β_0 which suggests that participants do not expect homogeneous and heterogeneous teams to be different, after controlling for beliefs about differences in individual productivity. We find $\beta_1 > 0$ which suggests that differences in beliefs about individual productivity translate into differences in beliefs about team productivity. The estimate for β_2 is not significant indicating that the two groups do not have

[26] We reject the hypothesis that the median person believes the difference to be equal to the true difference (p = 0.000, WSR). This result also holds for each ethnic group separately (p = 0.000, WSR).

different beliefs about the type-specificity of the production function, after controlling for beliefs about individual productivity differences. In summary, beliefs about individual productivity differences across types explain differences across homogenous and heterogeneous teams. In addition, homogenous teams are not generally believed to outperform heterogeneous teams, and these beliefs are not different across ethnic type of participant.

B) Animus-driven and belief-driven prejudice matter

We now show that taking prejudice into account substantially improves predictions in NoInfo compared to the benchmark case of accurate statistical discrimination (ASD). We find that ASD predicts discrimination rates inaccurately and that there is a substantial gap between decision makers' predicted earnings according to ASD and observed earnings. Taking the two types of prejudice into account provides much more accurate predictions for discrimination choices and explains the earnings gap almost entirely (97.2%) for one ethnic type and about half (48.5%) of the gap for the other type.

To show that prejudice matters for discrimination, we compare 4 scenarios which differ by assumptions about decision makers' animus and beliefs.

a) No animus, rational beliefs. ASD assumes that decision makers have no animus and maximize expected earnings given rational beliefs about average productivity of ethnic types.[27] Our experiment provides a rare opportunity to test the predictions of ASD because we can retrieve rational beliefs from the distribution of workers' output in phase 1 as follows. For each decision maker *i*, we sample observed round 1 output of two candidates of different types. We estimate the marginal product of labor (MPL) for *i* with either candidate using model A from table 2. Decision maker *i*'s price of choosing one worker over the other is the difference between these MPLs. By repeatedly sampling and averaging, we obtain the expected price for *i* of choosing one type over the other (see Appendix I for details). Because workers with Danish-sounding names are on average more productive than those with Muslim-sounding names in our sample, we find that ASD predicts that all decision makers choose the worker with the Danish-sounding name. However, only about half of decision

[27] While ASD maximizes expected earnings absent precise information about candidates' productivities, it does not yield the first-best outcome. Losses occur when choosing the candidate of the more productive type because decision makers by chance pick a less productive worker when type-productivity distributions overlap. The loss due to limited information is 2.5 percent of round 2 earnings. Yet, there is a clear incentive for ASD in NoInfo. In fact, earnings are 2.5 percent higher with ASD than with random choice of partner.

makers (49%) do so. The misprediction is particularly pronounced for decision makers with Muslim-sounding names (only 10.5 percent choose *other*).[28]

b) No animus, biased beliefs. This scenario serves to evaluate the predictive power of statistical discrimination with elicited (i.e. inaccurate) beliefs. We use the same procedure as in a) to retrieve elicited beliefs except that we draw from the distribution of elicited beliefs about round 1 output. We find that statistical discrimination *cum* biased beliefs does not improve predictions compared to ASD. Section A above has shown that elicited beliefs are quantitatively biased in the sense that the true productivity differences across types are underestimated. However, because beliefs were not strongly biased, belief-driven prejudice yields the same predictions as ASD.[29]

c) Animus, rational beliefs. This scenario serves to evaluate the predictive power of animus given rational beliefs. To calculate predictions, we use rational beliefs as described in a) and feed those beliefs into our estimate of taste-based discrimination treatment Info (see model 1 in table 3) to estimate the probability that decision maker *i* chooses *same*. By doing so, we assume that the distribution of animus-driven prejudice is the same in treatment Info and NoInfo. This is assumption is warranted since decision makers were randomly allocated to treatments.

Taking animus-driven prejudice into account improves the prediction for the decision makers with Danish-sounding names from 100 to 79.1 percent. This prediction is not statistically different from the observed 66.7 percent (p = 0.711, Fisher exact test).[30] The prediction for the decision makers with Muslim-sounding names is also improved. Now, 57.3 percent (rather than 100 percent) are predicted to choose *other*. Yet, the prediction is still different from the observed 10.5 percent (p = 0.013, Fisher exact test).

d) Animus, biased beliefs. In this scenario, we feed elicited beliefs (as described in b above) into our estimate of animus from treatment Info (as described in c above). Note that while

---

[28] The fact that the vast majority (89.5 percent) of decision makers with Muslim-sounding names chooses *same* in NoInfo suggests that any preference for a weekday that may have been present is swamped by ethnic preferences in our sample. Recall from section 3 that our randomized matching procedure guarantees that candidates are randomly allocated to weekdays.

[29] Note that this is the case in our experiment because discrimination choices are discrete. Had discrimination involved a continuous variable like wages, any bias in expected MPL would translate into a cost.

[30] Tests in this section assume an equal number of observations for predicted and observed discrimination rates.

biased beliefs do not make a difference for predictions given that decision makers have no animus in our design, they do make a difference given animus-driven prejudice. The reason is that the prediction moves discretely with beliefs absent animus (all choose the type believed to be more productive on average) but moves continuously in the presence of animus (the demand for discrimination is smooth, see figure 2).

We find that taking both types of prejudice into account further improves the predictions. The prediction is now perfectly accurate for decision makers with Danish-sounding names (67.3 vs. 66.7 percent observed).[31] The prediction also improves for decision makers with Muslim-sounding names, but there is still a some discrepancy (60.8 vs. 89.5 percent). However, the predicted and observed discrimination rates are not significantly different after accounting for prejudice ($p = 0.232$, Fisher exact test).

Table 4 shows how the gap between earnings with ASD and observed earnings can be explained by prejudice using the scenarios described above. The table shows earnings foregone to decision makers by deviating from ASD, in percent of decision makers' round 2 earnings with ASD. The total gap is 3.6 percent (or about €2.3 per decision maker). The gap is smaller for decision makers with Danish-sounding names (1.6 vs. 5.8 percent) because they tend to choose the Danish-sounding, i.e. on average more productive, candidate more often.

The top row of table 4 shows that statistical discrimination *cum* biased beliefs (scenario b) cannot account for the earnings gap. The second row shows the explanatory power of scenario c. We find that animus *cum* rational beliefs predicts a loss of 1.7 percent relative to ASD. Note that the predictions are rather different for the two ethnic types. Decision makers with Muslim-sounding names have higher losses (2.3 vs. 1.1 percent). The main reason for this difference is that our estimate of animus predicts that decision makers of either type choose *same* more often than *other*. Hence, decision makers with Danish-sounding names tend to choose the more productive type more often than the decision makers with Muslim-sounding names. Assuming animus *cum* rational expectations explains about 40 percent (= 2.3/5.8) and two thirds (= 1.1/1.6) of the gap for Danish-sounding and Muslim-sounding decision makers, respectively.

---

[31] This highly precise prediction is remarkable given its out-of-sample nature. Recall that the demand for taste-based discrimination is estimated by forcing all cost to be positive in Info while the (average) cost of discrimination is negative for decision makers with Danish-sounding names in NoInfo.

Table 4: Earnings foregone relative to earnings with accurate statistical discrimination (ASD)

Type of prejudice	Belief	Animus	Danish-sounding	Muslim-sounding	Overall
Belief-driven	Elicited	None	0.0	0.0	0.0
Animus-driven	Rational	Elicited	-1.1	-2.3	-1.7
Both	Elicited	Elicited	-1.5	-2.8	-2.2
Observed	-	-	-1.6	-5.8	-3.6

Notes: The table shows earnings foregone relative to the benchmark of accurate statistical discrimination in percent of decision makers' round 2 earnings. Rational beliefs are retrieved for each decision maker *i* by repeatedly sampling from candidates' observed round 1 output. We then estimate the marginal product of labor (MPL) using model A in table 2 for each draw. Elicited beliefs are retrieved analogously by drawing from elicited beliefs (see section 4.3). In row 1, *i* chooses the candidate of the type with the higher average MPL given elicited beliefs. In rows 2 and 3, we estimate probabilities of choosing *same* from model 1 in table 3 and using the average price according to rational or elicited beliefs, respectively. We use these probabilities to calculate a weighted average of earnings for either type.

The third row of table 4 shows the explanatory power of scenario d. We find that the loss predicted by both types of prejudice is about 2.2 percent of earnings in the benchmark case.[32] Note that biased beliefs do matter given animus (about half a percentage point). Thus, adding biased beliefs to animus-driven prejudice explains an additional 14 to 33 percent of the gap.

In summary, we find that accurate statistical discrimination (ASD) cannot explain observed outcomes. We find that both types of prejudice together explain about 60 percent of the gap between earnings with ASD and observed earnings. The gap is almost perfectly (97.2 percent) explained for decision makers with Danish-sounding names. For decision makers

[32] Note that earnings foregone in NoInfo refer to all decision makers. In contrast, the earnings foregone reported in Info (8.4 percent) refer to the average discriminator. The comparable number for all decision makers is a loss of 2.8 percent. The difference (2.8 vs. 2.2) is mainly due to the fact that the price of choosing *same* was much smaller in NoInfo than in Info (averages are 1.1 vs. 9.8 percent). This is the case for two reasons. First, the price is positive by design in Info while it is positive or negative in NoInfo, depending on the type of decision maker. Second, in NoInfo the average price is relevant for choices while in Info it is the realization of a random draw, and some of these have high values. A decision maker with a strong animus discriminates in both treatments, but the implied price paid for this animus-driven prejudice is lower in NoInfo than Info.

with Muslim-sounding names, prejudice provides a much better prediction than ASD (48.2 percent of the gap is explained), but a considerable unexplained gap remains.[33]

5 Concluding remarks

This study develops a novel experimental approach to measuring the price of ethnic prejudice paid by discriminators in the work place. We show that part of this price is paid deliberately and is due to animus, and part of the price is paid unintentionally and is due to biased beliefs.

We find that statistical discrimination along with rational expectations (i.e. accurate statistical discrimination) grossly mispredicts observed behavior. Compared to this benchmark, decision makers leave about 4 percent of earnings on the table. We show that about 60 percent of this earnings gap can be accounted for by animus and belief-driven prejudice. We isolate taste-based discrimination by controlling for beliefs, i.e. by informing decision makers about the true productivity of job candidates, and by randomly assigning a price of discrimination to decision makers. Using a sample from Denmark, we find that discrimination is common even at a substantial price, that majority and minority groups are equally likely to discriminate for given prices, and that the demand for discrimination is remarkably elastic. Our best estimate is that the probability to discriminate falls by about 9 percent if the price of discrimination goes up by 10 percent. We use this estimate together with elicited beliefs to evaluate the predictive power of animus-driven and belief-driven prejudice.

Below, we discuss three potential sources of mismeasurement of prejudice due to selection effects and conclude that selection is not likely to have caused bias in one way or another. Finally, we emphasize that our quantitative findings should not be extrapolated to employment decisions in large firms without further consideration because incentives for personnel managers may be opaque or differ substantially from the sharp and controlled incentives for decision makers in our experiment.

First, we may underestimate animus in the general population because our sample is not representative of the Danish population. We recruit juveniles from secondary schools in

[33] We can only speculate what may explain the remaining gap. A possibility is "implicit discrimination" (Bertrand et al. 2005). However, it is not entirely clear why implicit discrimination should be more important for minority than for majority decision makers.

Copenhagen who have very similar age and education and are all fluent in the majority language. Such relatively well-educated and integrated juveniles as a group may have systematically lower animus than the average Dane or Muslim living in Denmark. In fact, available research suggests that (voiced) animus decreases with education and income but increases with age (e.g. Charles and Guryan 2008).

Second, we may over- or underestimate differences in animus across ethnic types. We find that minority and majority groups are equally likely to discriminate for a given price. This result is surprising in the light of evidence suggesting that minorities have more pronounced "homophily" (in the diction of Curarrini et al. 2009) than majorities. We may underestimate the difference due to unobserved heterogeneity in income in our sample. While the evidence presented in Charles and Guryan (2008) suggests that animus decreases with income, taste-based discrimination may well also increase with income (if it is a "normal" good). However, we may overestimate the difference due to a subtle name-related selection effect. A juvenile is classified as having a Muslim-sounding name in our experiment if his parents chose such a name, but is classified as having a Danish-sounding (or other) name if they did not. If the name choice by parents is correlated with animus, we would tend to overestimate differences in animus across ethnic groups. However, this effect seems to be of minor relevance since we find no difference in animus across ethnic types.

Third, we may over- or underestimate the relevance of belief-driven prejudice. We elicit beliefs on a different sample of juveniles and argue that elicited beliefs are a precise proxy of decision makers' beliefs in NoInfo. This claim seems plausible because both groups are similar in observables, both groups have an incentive to form beliefs, and, perhaps most importantly, we find that elicited beliefs provide a more precise prediction (given animus) of observed behavior than rational beliefs. Yet, elicited beliefs may be more or less accurate than decision makers' beliefs in NoInfo. On the one hand, beliefs may be less accurate because participants in the elicitation study are not experienced in the work task. On the other hand, elicited beliefs may be more accurate because participants were given explicit incentives for guessing correctly and may have thought more explicitly about how others perform.

A remarkable side result of our study is that, after controlling for individual productivity differences, minority and majority workers are equally productive in teamwork whether they work with someone from the same or the other ethnic type. In addition, we find no evidence for the claim that majority workers think they cannot work equally well with a minority worker than with a majority worker, all else equal. Thus, this type of belief-based prejudice

about teamwork receives no support in our study. However, we do find evidence for a different type of belief-based prejudice. We find that participants of both types underestimate the remarkably pronounced differences in productivity across types. Thus, we find that both majority and minority types seem, perhaps surprisingly, to expect less of a difference in productivity that there in fact is.

The extent to which the quantitative estimates from our experiment extrapolate to hiring choices, in particular in large firms, must remain an open issue for two reasons. First, we may over- or underestimate the importance of belief-driven prejudice compared to personnel managers who may have more or less accurate beliefs than decision makers our sample. On the one hand, personnel managers in large firms may be able to draw on extensive internal statistics and therefore have more accurate beliefs about the average productivity by ethnic type than decision makers in NoInfo. On the other hand, the work task in our experiment was well-defined and simple compared to collaborative tasks in large firms. It is therefore relatively easy for our decision makers to predict productivity accurately. Second, we may over- or underestimate the sensitivity of taste-based discrimination to the price of prejudice because decision makers (in Info) faced a clear and known price for discrimination while incentives may be opaque or weak for a personnel officer in a large firm. Decision makers in our experiment are directly affected (monetarily and non-monetarily) by their choices because they make a consequential choice of whom to work with in a team. In contrast, personnel managers do not necessarily physically work with new hires and may also be largely shielded from monetary consequences of their choices. On the other hand, large corporations may have particular policies (like affirmative action programs) on discrimination in place which may provide incentives against discrimination.

In conclusion, our results suggest that belief-driven and animus-driven ethnic prejudice are important causes of ethnic discrimination in the workplace, and need to be taken into account above and beyond the theory of accurate statistical discrimination.

References

Altonji, J.G. and Blank, R. (1999): Race and Gender in the Labor Market. *Handbook of Labor Economics* 3(13): 3143-3259.

Altonji, J.G. and Pierret, C.R. (2001): Employer Learning and Statistical Discrimination. *Quarterly Journal of Economics* 116(1): 313-50.

Anderson, L.R., Fryer, R.G. and Holt, C.A. (2006): Discrimination: Experimental Evidence from Psychology and Economics, in: W.M. Rodgers (ed.): *Handbook on the Economics of Discrimination*. Cheltenham: Edward Elgar, 97-115.

Arrow, K.J. (1973): The Theory of Discrimination. In: O. Ashenfelter and A. Rees (eds.): *Discrimination in Labor Markets*. Princeton, N.J., Princeton University Press.

Ayres, I. and Siegelman, P. (1995): Race and Gender Discrimination in Bargaining for a New Car. *American Economic Review* 85(3): 304–321.

Baccara, M., Collard-Wexler, A., Felli, L. and Yariv, L. (2009): Gender and Racial Biases: Evidence from Child Adoption. Working paper NYU.

Becker, G.S. (1957): *The Economics of Discrimination*. University of Chicago Press.

Bertrand, M., Chugh, D. and Mullainathan, S. (2005): Implicit Discrimination. *American Economic Review* 95(2): 94-98.

Bertrand, M. and Mullainathan, S. (2004): Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review* 94(4): 991-1013.

Carlsson, M. and Rooth, D.O. (2007): Evidence of Ethnic Discrimination in the Swedish Labor Market Using Experimental Data. *Labour Economics* 14: 716-729.

Caruso, E.M., Rahnev, D.A. and Banaji, M.R. (2009): Using Conjoint Analysis to Detect Discrimination: Revealing Covert Preferences from Overt Choices. *Social Cognition* 27(1): 128-37.

Charles, K.K. and Guryan, J. (2008): Prejudice and Wages: An Empirical Assessment of Becker's "The Economics of Discrimination". *Journal of Political Economy* 166(5): 773-809.

Currarini, S., Jackson, M.O. and Pin, P. (2009): An Economic Model of Friendship: Homophily, Minorities, and Segregation. *Econometrica* 77(4): 1003-1045.

Daniel, W. (1968): *Racial Discrimination in England*, Middlesex: Penguin Books.

Fershtman, C. and Gneezy, U. (2001): Discrimination in a Segmented Society: An Experimental Approach. *Quarterly Journal of Economics* 116(1): 351-77.

Goldin, C. and Rouse, C. (2000): Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians. *American Economic Review* 90(4): 715-741.

Hamermesh, D.S. and Biddle, J.E. (1994): Beauty and the Labor Market. *American Economic Review* 84(5): 1174-94.

Heckman, J.J. (1998): Detecting Discrimination. *Journal of Economic Perspectives* 12(2): 101-116.

Heckman, J.J. and Siegelman, P. (1993): The Urban Institute Audit Studies: Their Methods and Findings. In M. Fix and R. Struyk, eds. *Clear and Convincing Evidence: Measurement of Discrimination in America*, 187-258.

Holm, H.J. (2001): What's in a Name? An Ethnical Discrimination Experiment. Working paper, Lund University.

Jowell, R. and Prescott-Clarke, P. (1970): Racial Discrimination and White-collar Workers in Britain. *Race* 11: 397-417.

Kawakami, K., Dunn, E., Karmali, F. and Dovidio, J.F. (2009): Mispredicting Affective and Behavioral Responses to Racism. *Science* 323: 276-278.

Lang, K. (1986): A Language Theory of Discrimination. *Quarterly Journal of Economics* 101(2): 363-382.

Levitt, S. (2004): Testing Theories of Discrimination: Evidence from Weakest Link. *Journal of Law and Economics* 47: 431-452.

List, J. (2004): The Nature and Extent of Discrimination in the Marketplace: Evidence from the Field. *Quarterly Journal of Economics* 119(1): 49-89.

List, J. (2006): Field Experiments: A Bridge Between Lab and Naturally Occurring Data. *Advances in Economic Analysis and Policy* 6(2): Article 8.

Möbius, M.M. and Rosenblat, T.S. (2006): Why Beauty Matters. *American Economic Review* 96(1): 222-235.

Neumark, D. (2010): Detecting Discrimination in Audit and Correspondence Studies. NBER working paper 16448.

Neumark, D., Bank, R.J. and van Nort, K.D. (1996): Sex Discrimination in Restaurant Hiring: an Audit Study. *Quarterly Journal of Economics* 111(3): 915-941.

Oreopoulos, P. (2009): Why Do Skilled Immigrants Struggle in the Labor Market? A Field Experiment with Six Thousand Resumes. NBER Working Paper 15036.

Pager, D. (2007): The Use of Field Experiments for Studies of Employment Discrimination: Contributions, Critiques, and Directions for the Future. *Annals of the American Academy of Political and Social Science* 609: 104-133.

Pager, D., Western, B. and Bonikowski, B. (2009): Discrimination in a Low-Wage Labor Market. *American Sociological Review* 74(5): 777-799.

Phelps, E. (1972): The Statistical Theory of Racism and Sexism. *American Economic Review* 62(4): 659-661.

Pope, D.G. and Sydnor, J.R. (2011): What's in a Picture? Evidence of Discrimination from Prosper.com. *Journal of Human Resources* 46(1): 53-92.

Riach, P.A. and Rich, J. (2002): Field Experiments of Discrimination in the Market Place. *Economic Journal* 112(483): 480-518.

Yinger, J. (1998): Evidence on Discrimination in Consumer Markets. *Journal of Economic Perspectives* 12(2): 23-40.

Zizzo, D. (2010): Experimenter Demand Effects in Economic Experiments. *Experimental Economics* 13(1): 75-98.

Appendices

Appendix A: Flyer used for recruiting

Appendix B: Location and participants

Figure B1: Location of schools from which participants were recruited
Figure B2: The control room
Figure B3: Floor plan

Appendix C: Description of the work task

Figure C1: Photograph of a workstation

Appendix D: Validation of classification of first names

Table D1: Effectiveness of first names as marker of ethnic type

Appendix E: Using productivity differences as proxy for the price of discrimination

Table E1: The demand for discrimination using output differences as a proxy for *Price*

Appendix F: Robustness of price effect with respect to the decision maker's productivity

Table F1: Discrimination and the decision maker's productivity

Appendix G: Testing for random assignment of price (simulation)

Appendix H: Eliciting productivity beliefs

Table H1: Average output guesses by participants in complementary study

Appendix I: Decomposition of the earnings gap

Appendix A: Flyer used for recruiting

Tjen penge!

Har du lyst til at tjene ekstra penge?

Københavns Universitet skal sende 40.000 invitationer til vores nye internet platform (www.econ.ku.dk/ILEE), og vi har brug for hjælp til at pakke brevene.

Du skal kunne arbejde 2 gange 2 timer. De første 2 timer skal være i uge 49 (3. - 7. dec.), og de sidste 2 timer i uge 50/51 (10. - 19. dec.). Arbejdet foregår i centrum af København og vi tilbyder en god løn.

Arbejdstiden vil kunne være hverdage mellem kl. 13 og kl. 21. Jo mere fleksibel du er, desto større chance er der, for at vi kan bruge dig. Vi aftaler naturligvis det specifikke tidspunkt i god tid inden arbejdet.

Du vil blive aflønnet efter, hvor mange breve du pakker, og vi forventer i gennemsnit at betale cirka 180 kr./time.

Hvis du er interesseret så ring på tlf. 35 32 44 04 / 35 32 30 59 mellem klokken 10 og 18 eller send en e-mail med navn og telefonnummer til ILEE@econ.ku.dk.

KØBENHAVNS UNIVERSITET

Translation: Earn money! Would you like to earn some extra money? The University of Copenhagen has to mail 40'000 invitation letters for a new internet platform, and we are looking for help to pack these letters.
You are supposed to work twice for 2 hours. The first 2 hours are in week 49 … the second in week 50/51.
Work is to be done in the city center and we pay a good salary. Work times are between 1 p.m. and 9 p.m. You are more likely to be hired if you are more flexible with respect to work times. We will of course make a specific agreement with sufficient notice.
You will be paid according to how many letters you pack and we expect to pay about 180 kr. (about €24) per hour.
Call us on …between .. and .. or send an e-mail with your name on phone number to … if you are interested.

Appendix B: Location and participants

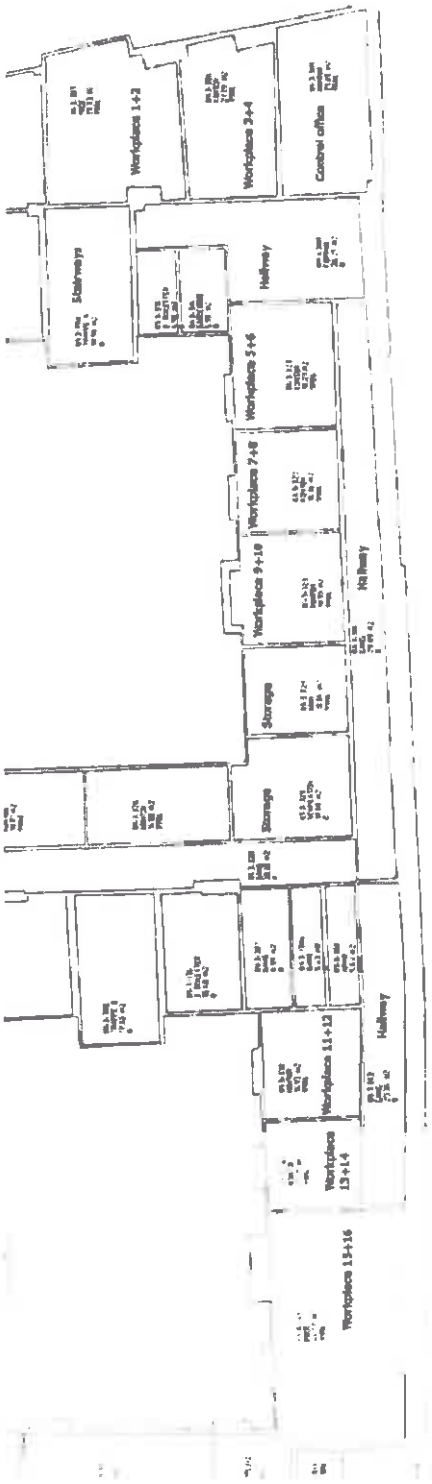
Figure B1 shows the secondary schools from which participants were recruited for the experiment (red symbols), for the belief elicitation and name validation studies (blue symbols) and the pre-test (purple marker in the lower left corner). The flag indicates the location of the University premises where work was carried out.

Figure B1: Location of schools from which participants were recruited[34]

[34] The figure has more than eleven red markers as some of the schools where we recruited for the experiment have several campuses in Copenhagen.

***Figure B2*:** The control room

Figure B3: Floor plan

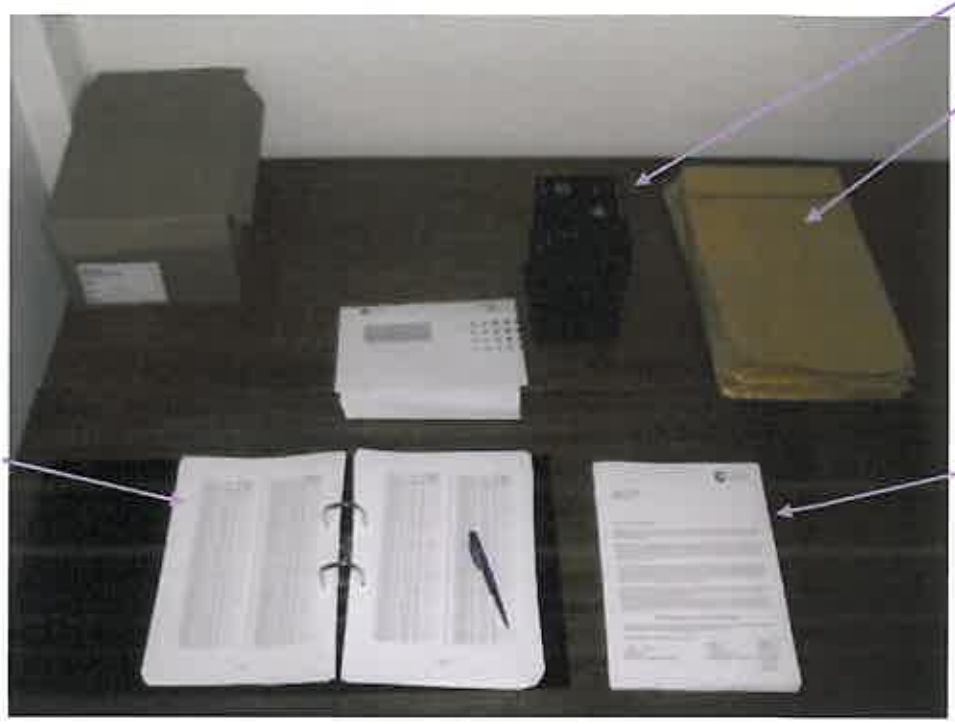


The University of Copenhagen generously provided us with an entire floor (app. 320 m^2) of 11 offices which were furnished with tables and chairs. Two offices were used for storage of materials, one office was used as control room (see figure B2) and work was carried out in the remaining eight offices.

Appendix C: Description of the work task

The participants were seated in a two-person office at a workstation facing the wall. Figure C1 shows a photograph of the workstation.

Figure C1: Photograph of a workstation



Each letter had a an ID number (ranging from 12,000 to 51,999). The order of the letters was randomized so each participant could have letters from the entire interval.

The 40,000 letters had to be sorted into 5 main categories (A to E). These were then split further into a total of 96 subcategories (A-1 to E-96). The sub-categories were assigned randomly and were not printed on the letters. Each participant would get letters belonging to six subcategories and would have to sort the letters accordingly.

For each letter, the task was to: Look up the letter's ID number in a binder with 600 pages and see which category (A-1 to E-96) the letter belongs to; Look up the category type (A to E) in a separate list and see whether the letter should include a gift (letters in categories B and D should include a small foam puzzle) ; Fold the letter and stuff it into an envelope. If category B or D, then also include a gift ; Close the envelope ; Sort the envelope into the collection envelope with the subcategory label written on the outside.

The participants received both oral and written instructions on how to do the task. These instructions were given individually and we demonstrate how to pack a letter. The participant then packed a under supervision to verify understanding of the procedure. If successful, the participants worked alone for 90 minutes. An alarm clock was set in the control room to enforce to time limit. After the 90 minutes, we stopped the participants and counted the number of letters packed. In total, the participants spend less than two hours at the University in each round.

Appendix D: Validation of classification of first names

As in correspondence tests, we use names as a marker of ethnicity. However, we do not use fictitious and highly stereotypical names but the actual names of workers. We categorize these names into ethnic types using our judgment complemented by lists of "typical" Danish and Muslim names we found on the web (such as www.muslimbabynames.net).

To test if actual names are effective markers of ethnicity, we run a complementary study with n = 144 juveniles in a secondary school on the outskirts of Copenhagen where we do not recruit for the experiment. The questionnaire (available from the authors on request) presents respondents with 4 randomly drawn pairs of candidates (i.e. using the actual names and actual pairs decision makers faced) and asks them classify the names as either Danish or Muslim. More specifically, respondents have the option to classify either, both or none of the two names as 'Danish' or 'Arab/Muslim'. We randomize the order of names for a given choice in any given pair. This task is presented to respondents as part of a "classification study" which also contains 9 other, unrelated, tasks (e.g. classify cities as German or French). Participants are paid a flat fee of DKK 100 (€13.3) for completing the survey.

Table D1: Effectiveness of first names as marker of ethnic type

	Boys		Girls		
Concordance	Danish-sounding	Muslim-sounding	Danish-sounding	Muslim-sounding	*Overall*
Danish names	80%	87%	84%	94%	*83%*
Muslim names	97%	94%	86%	92%	*92%*
Overall	*89%*	*90%*	*85%*	*93%*	*88%*
Confound					
Danish names	2%	3%	3%	5%	*2%*
Muslim names	0%	0%	1%	0%	*1%*
Overall	*1%*	*2%*	*2%*	*3%*	*1%*

Notes: The table shows the percentage (over of all names and respondents) of classifications in the survey study that are in line ("concordance") or conflict ("confound") with the classification into ethnic types in the experiment. Concordance occurs, for example, if a name we classify as Danish-sounding in the experiment is classified by respondents as Danish-sounding. Confound occurs, for example, if a name we classify as Danish-sounding is classified by respondents as Muslim-sounding. The number of respondents is n = 144.

Table D1 shows that concordance rates are very high and confound is rare. In particular, the last column shows that 83 percent of the names we classify as Danish-sounding and 92 percent of those we classify as Muslim-sounding are categorized by respondents in concordance with our classification. Importantly, it very rarely happens (1 percent of the cases) that names we classify as belonging to one ethnic type are classified as belonging to the other category by respondents. Concordance and confound rates are similar for respondents with Danish-sounding and Muslim-sounding names.

Appendix E: Using productivity differences as proxy for the price of discrimination

This appendix shows that our main result in Info (that an increase in price causally reduces taste-based discrimination) is robust to using a different type of team production function to estimate prices.

In section 4.1, we estimate the price from the marginal productivity of labor obtained from a particular type team production function (model A in table 3). We then use these (randomly assigned) prices to estimate the demand for discrimination (and the willingness to pay). By doing so, we assume that the price, and implicitly also the team production function, is known to decision makers. To demonstrate robustness, we use "raw" round 1 output differences as a proxy for the price in the estimation of the demand for discrimination and therefore tie the price of prejudice directly to observables. We find very similar results either way.

Table E1 replicates the analysis in table 3 using (half of) the difference of round 1 output between the candidates as a proxy for the price of discrimination. The coefficient of $\Delta Prod_{jk}$ in model (8) shows that if price goes up by €1, decision makers are about 3 percent less likely to discriminate. This estimate is similar to our result for *Price* in table 3 (3.0 vs. 3.6 percent). Also note that models (9) to (11) yield very similar results as models (2) to (4) in table 3.

Table E1: The demand for discrimination using output differences as a proxy for *Price*

Dependent variable: Discr	(8)	(9)	(10)	(11)
$\Delta Prod$	-0.030**	-0.029**	-0.028**	-0.029*
	(0.013)	(0.014)	(0.014)	(0.016)
Danish-sounding		0.014		0.088
		(0.160)		(0.273)
Male		-0.063		-0.138
		(0.152)		(0.266)
Danish-sounding * $\Delta Prod$			-0.001	-0.010
			(0.020)	(0.035)
Male * $\Delta Prod$			-0.005	0.010
			(0.017)	(0.029)
N	37	37	37	37
Adj. R^2	0.073	0.076	0.074	0.079

Notes: The table shows average marginal effects for probit regressions. Numbers in parentheses are robust standard errors. The dependent variable *Discr* = 1 for a discriminator and 0 otherwise. The variable $\Delta Prod_{jk}$ is the difference in output in round 1 by *other* minus output by *same*. To make the numbers comparable, we multiply the difference by 0.5 as the joint output was split among the two team members and express values in Euros, i.e. multiply with €0.5 per letter packed. *Danish-sounding* and *Male* are dummy variables characterizing decision maker *i*. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Appendix F: Robustness of price effect with respect to the decision maker's productivity

Our discussion of the response of taste-based discrimination to the price of prejudice in Info in section 4.1 is entirely cast in terms of earnings foregone by choosing one candidate over the other, i.e. is based on opportunity cost. Below, we address issues relating to the absolute and relative productivity of the decision maker.

Table F1 investigates if decision makers with high productivity in round 1 tend to be less likely to discriminate. Such an effect is plausible if those with a strong preference for money work hard and also tend to choose a co-worker primarily on the basis of monetary concerns. But we find that the effect is weak is best ($Prod_1$ is insignificant in models 5 and 6). The table also serves to investigate whether the decision maker's productivity in round 1 relative to the productivities of the two candidates biases our estimates of the demand for discrimination. Our conclusion from the discussion below is that it does not.

Table F1: Discrimination and the decision maker's productivity

Dependent variable: Discr	(5)	(6)	(7)
Price	-0.030*	-0.030**	-0.017
	(0.016)	(0.015)	(0.018)
$Prod_1$	-0.046*	-0.044	-0.043
	(0.026)	(0.033)	(0.032)
$Prod_1^2$	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)
Abs. distance to *same*		0.001	-0.003
		(0.008)	(0.007)
Same candidate below			-0.153
			(0.199)
Both candidates below			0.101
			(0.268)
N	37	37	37
R^2	0.147	0.147	0.177

Notes: The table shows average marginal effects estimated from Probit regressions. Numbers in parentheses are robust standard errors. The dependent variable *Discr* = 1 for a discriminator and 0 otherwise. The variable *Price* is expressed in Euros. $Prod_1$ and $Prod_1^2$ are decision maker *i*'s productivity and its square in round 1. *Abs. distance to "same"* is the absolute difference in round 1 productivity between decision maker *i* and the candidate of the same ethnic type as *i*. *"Same" candidate below* is a dummy variable taking the value 1 if the productivity of the decision maker in round 1 is between the two candidates. *Both candidates below* is a dummy variable taking the value 1 if the productivity of the decision maker in round 1 is higher than that of both candidates. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A potential concern with using an opportunity cost concept is that it does not take relative standing into account. Due to random matching of decision makers into triples, decision makers have a choice between candidates who can be more or less similar to the decision

maker in terms of round 1 productivity. A particular concern is that choosing *same* may not reflect a preference for an ethnic type, but a preference for a co-worker with similar productivity. For example, a decision maker may choose *same* to avoid peer pressure and feeling uncomfortable when working with a much more productive co-worker. Model (6) in table F1 includes a variable *Abs. distance to "same"* which measures the absolute productivity difference between the decision maker and the candidate of the same ethnic type. The insignificant coefficient suggests that this concern does not affect the choice of co-worker.[35]

Model (7) in table F1 investigates a potential confound of loss aversion and taste-based discrimination. Due to the randomness of our matching procedure, decision makers have a choice between a) two candidates which are both less productive, b) both more productive, or c) a more and a less productive candidate. Compared to the case of being in a team with a co-worker with the same productivity, discrimination in case a) means incurring an additional loss, in b) foregoing an additional gain, and in c) incurring a loss rather than making a gain. Thus, loss aversion predicts that choosing *same* is less likely in case c) than in a) or b), and less likely in a) than b) for a given price of discrimination. To test, we add *"Same" candidate below* (equal to 1 in case c) and *Both candidates below* (a dummy equal to 1 in case a). The insignificant estimates suggest that loss aversion does not seem to have affected the choice of a co-worker. However, this result should be taken with a grain of salt due to multi-collinearity and the large number of explanatory variables compared to the number of observations.

[35] We also find that decision makers do not have a bias in favor of the candidate with more "similar" productivity in a simple non-parametric test. Out of 37 decision makers, 21 choose the "closer", 16 the "further" candidate. This split is not statistically different from a 50:50 split ($p = .560$, χ^2 test).

Appendix G: Testing for random assignment of price (simulation)

A precondition for identifying the causal effect of prices on discrimination choices in treatment Info is that the price of discrimination (i.e. the opportunity cost choosing *same* over *other*) is randomly assigned to decision makers. In particular, the distribution of animus and the distribution of the prices must be independent.

Our matching procedure (see section 3) is sequential and matches (randomly drawn) decision makers with candidates from a pool of suitable candidates. That is, once a decision maker is determined, the candidates are drawn from a constrained set (e.g. the candidates and the decision maker have to be available on the same days). A possible concern is that our matching procedure caused selection in the sense that characteristics of the decision maker constrain the set of set of suitable candidates in such a way that the resulting distribution of prices is not random and independent of decision makers' animus.

Below, we provide three tests for random assignment of prices to decision makers. The tests do not reject the hypothesis of random assignment.

First, we test if the distribution of prices observed in our experiment is normal. Unconstrained random drawing of pairs of candidates implies that the distribution of $Price_i$ follows (half a) normal distribution. Because $Price_i$ is positive by design in Info, we mirror the experimental distribution on 0, and test this distribution for normality using standard tests. We cannot reject the normality assumption ($p = 0.818$, Shapiro-Wilk; $p = 0.721$, Shapiro-Francia; $p = 0.901$, Skewness/Kurtosis test for normality).

Second, we test if the sequentiality of our matching procedure caused a bias in the distribution of productivity differences between candidates. We test for productivity differences because these are directly observable and are a good proxy to $Price_i$ (see appendix E for a discussion). In particular, we test if the observed distribution of productivity differences is different from a simulated distribution which is obtained from random draws without (unintended) constraints. The simulated productivity differences are obtained by sampling from all participants who complete round 1 ($n = 162$) with two constraints which are intended consequences of our design (rather than unintended consequences of sequential sampling). Our simulation imposes that a decision maker is always matched with candidates of the same gender (to avoid confound of gender and ethnicity) and that *same* is by design less productive than *other* (to make choices informative). We sample 1'000 productivity differences for each type of decision maker. From this pool, we randomly draw 37 productivity differences and test the resulting distribution against the experimentally observed distribution using Mann-Whitney (MW) and Kolmogorov-Smirnov (KS) tests. We repeat the draw and run the tests 1'000 times. At a level of significance α, we expect fewer than α percent of these tests to reject (i.e. to have a p-value $< \alpha$) if the null is true. At $\alpha = 0.05$, we find that these tests reject in less than 1 percent of the cases (MW: 0.009, KS: 0.005). At $\alpha = 0.1$, we find that the tests reject in less than 3 percent of the cases (MW: 0.029, KS: 0.009). In summary, our sequential matching procedure yields productivity differences which are indistinguishable from purely random

draws of candidates and the sequential matching we use does therefore not seem to bias prices.

Third, we test for the independence of the distribution of animus and the distribution of prices by means of a simulation. This is a joint test for independence and other assumptions which are simultaneously imposed in the simulation. In particular, the simulation imposes a normal distribution of prices, a normal distribution of animus (an assumption we make in using probit regressions), and independence of the two distributions. We also impose utility maximization in that the decision maker discriminates if and only if $a_i \geq Price_i$, just as we do in our estimations (see section 4.2). We compare the simulated distributions to the observed distribution in the experiment using non-parametric tests. We find that our experimental observation is likely to come from a population in which the assumptions above, including independence, jointly hold.

We proceed as follows. We randomly sample $n = 37$ pairs of $Price_i$ and a_i. $Price_i$ is drawn from the best fit of a normal distribution to estimated prices and a_i is drawn from the estimated distribution as explained in section 4.2B. If $a_i \geq Price_i$, we assign a value of $Discr_i = 1$, and = 0 otherwise. We calculate the conditional distribution of price for discriminators ($Discr_i = 1$) and non-discriminators, and the share of discriminators. We test these 3 distributions against the respective distributions as observed in the experiment using non-parametric tests. We repeat 1000 times for each distribution and expect a share of less than α (the significance level) of these tests to have p-values $< \alpha$ if the null hypothesis is true.

For the conditional distribution of the price of discriminators we find no significant difference between simulated and observed data. At $\alpha = 0.05$, we find that non-parametric tests reject in less than 3 percent of the cases (Mann-Whitney (MW): 0.024, Kolmogorov-Smirnov (KS): 0.020). At $\alpha = 0.1$, we find that the tests reject in 5 percent or less of the cases (MW: 0.050, KS: 0.040).

For the conditional distribution of the price of non-discriminators we find no significant difference between simulated and observed data. At $\alpha = 0.05$, we find that non-parametric tests reject in less than 2 percent of the cases (MW: 0.011, KS: 0.010). At $\alpha = 0.1$, the tests reject in 3 percent of the cases (MW: 0.030, KS: 0.030).

We find a mean simulated discrimination rate of 38.4 percent (observed is 37.8 percent, $n = 37$). We run 1'000 Chi-square tests to test for differences in the simulated and observed discrimination rate. At $\alpha = 0.05$, we find that the tests reject in less than 1 percent of the cases (χ^2: 0.007), at $\alpha = 0.1$, the tests reject in less than 2 percent of the cases (χ^2: 0.013).

In conclusion, the tests for the conditional prices of discriminators, of non-discriminators and the discrimination rates reveal that the observed data in our experiment does not look different from simulated data imposing random allocation of prices to decision makers.

Appendix H: Eliciting productivity beliefs

We recruit n = 353 juveniles to elicit beliefs about individual and team output across ethnic types in the letter packing task from two secondary schools where we do not recruit for the experiment. We carefully explain the work task to these participants and ask them to guess the productivity of actual workers in our experiment. We provide incentives for guessing correctly (the full questionnaire is available from the authors on request).

In particular, we present participants with a table of 7 randomly selected workers of the same gender and ask them to guess how many letters each worker packed when working in isolation in round 1. We also ask them to guess round 2 output for 6 randomly selected teams (2 homogeneous Danish-sounding, 2 homogeneous Muslim-sounding and 2 heterogeneous teams). As a point of reference, we provide participants with the observed median production in rounds 1 and 2. In total, 204 juveniles with Danish-sounding and 149 with Muslim-sounding names participate (42 have names that are classified as "other" and are omitted from the study). Beliefs are incentivized using a quadratic scoring rule. Average earnings are €13.6.

Table H1 shows that both types of participants tend to believe that workers with Danish-sounding names are more productive than workers with Muslim-sounding names when working alone (109 vs. 106 and 101 vs. 98, respectively). Remarkably, these beliefs about individual productivity differences across ethnic types are qualitatively in line with our results for round 1 production (116 vs. 100). However, both types of participants underestimate the true difference across ethnic types (3 vs. 16 letters).

Concerning team output, table H1 shows that both groups expect homogeneous Danish-sounding teams to be more productive than productive than heterogeneous teams which, in turn, are believed to be more productive homogeneous Muslim-sounding teams. The differences in beliefs about team production almost perfectly reflect the differences in beliefs about individual production. In particular, expected output increases by 3 letters by replacing a team worker with a Muslim-sounding name by one with a Danish-sounding name. Note that this almost perfect correspondence holds for participants of both ethnic types.

Table H1: Average output guesses by participants in complementary study

Participant	Individual workers		Teams		
	Danish-sounding	Muslim-sounding	Danish-sounding	Muslim-sounding	Hetero-geneous
Danish-sounding	109	106	225	220	223
Muslim-sounding	101	98	215	207	211

Notes: The table shows the average guesses for output of individuals and teams by participants in the belief elicitation study with Danish-sounding (n = 204) and Muslim-sounding (n = 149) names.

Appendix I: Decomposition of the earnings gap

This appendix describes how we decompose the earnings gap in treatment NoInfo into an animus-driven and a belief-driven component in section 4.3B. The earnings gap is the difference in decision makers' total earnings between the benchmark case of accurate statistical discrimination (ASD) and observed earnings. A gap results if decision makers choose a worker of the on average less productive type. Such a choice can result from holding a biased belief about the average price by type, from animus against a type of worker, or from other sources (unexplained part). ASD is profit-maximizing given available information and assumes that any prejudice is absent. That is, ASD assumes decision makers have rational beliefs on the price of discrimination and no animus.

Rational expectations ($Price_i^{RE}$) are determined for each i of the $n = 37$ decision makers as follows. We draw two co-workers (of the same gender as i) from the population of workers in our experiment (161 other workers, see table 1). We estimate team output with each drawn co-worker using i's production in round 1 and model A in table 2. The price of discrimination is then the difference in i's estimated earnings with either type. We repeat this procedure 1'000 times to obtain a distribution of $Price_i^{RE}$.

Elicited expectations ($Price_i^{EE}$) are determined in the same way as in the case of rational expectations except that we do not draw from the true distribution of round 1 output but from the distribution of elicited beliefs about round 1 output. Beliefs are elicited for 353 participants in the belief elicitation study (see section 4.3).

We use the means of these distributions (μ_i^{RE} and μ_i^{EE}, respectively) to predict behavior for i in 4 scenarios which differ by expectations formation (rational vs. elicited) and animus (no vs. as measured in treatment Info). The difference between the benchmark case of ASD and observed outcomes is decomposed into an animus-driven and a belief-driven component (see also section 4.3).

Absent any animus and assuming rational expectations, i chooses *same* if $\mu_i^{RE} < 0$ and *other* otherwise. The case is analogous for elicited expectations and no animus: i chooses *same* if $\mu_i^{EE} < 0$ and *other* otherwise. Note that as long as μ_i^{RE} and μ_i^{EE} have the same sign, they yield the same prediction for the choice of partner. In particular, we find that $\mu_i^{RE} < 0$ and $\mu_i^{EE} < 0$ for all decision makers with Danish-sounding names, and $\mu_i^{RE} > 0$ and $\mu_i^{EE} > 0$ for all decision makers with Muslim-sounding names.

To predict behavior in the case with animus, we feed μ_i^{RE} and μ_i^{EE} into model 1 from table 3 to calculate the probability that i chooses the co-worker of the same ethnic type ($Prob_i^{RE}$ and $Prob_i^{EE}$). We use these probabilities to calculate expected earnings and report earnings foregone by ethnic type from deviating from ASD in each scenario in table 4. Note that because $\mu_i^{RE} \neq \mu_i^{EE}$, and because our estimate of the demand for discrimination is continous (see figure 2), taking biased beliefs into account changes predictions for both ethnic types given animus-based prejudice in table 4.

Chapter 2

Correlates and Consequences of Distributional Preferences: an Internet Experiment

Morten Hedegaard

Correlates and Consequences of Distributional Preferences: an Internet Experiment

Morten Hedegaard*

March 2011

We investigate the correlates and consequences of distributional preferences in an experiment carried out over the internet with a large, heterogeneous subject pool. First, we find substantial heterogeneity in the distribution of distributional preferences and, in line with previous literature, we find that efficiency maximization seems to be more important than inequality aversion. Second, we find that gender, age, subjects' expectation of being treated fairly, cognitive reflection and IQ correlate with distributional preferences. Third, we investigate the link between distributional preferences and contributions in the standard public good game. We find that subjects who are efficiency maximizers, inequality averse and have maximin preferences contribute more than those who are selfish, even after controlling for beliefs. Finally, we find that taking distributional preferences into account explains almost half of the difference between observed behavior in the public good game and the prediction of standard economic theory.

Keywords: social preferences, distributional preferences, internet experiment, corporation, public good

JEL Classification numbers: C72, C91, D64

---

* University of Copenhagen, Department of Economics, Øster Farimagsgade 5, building 26, DK-1353 Copenhagen K, Denmark. Morten.Hedegaard@econ.ku.dk.

I thank Jean-Robert Tyran and Rudolf Kerschbamer for useful comments. I thank Erik Wengström and the rest of the iLEE team for providing data from the first wave of internet experiments. I also thank Eva Gregersen, Nikolaos Korfiatis and Thomas Alexander Stephens for their support in conducting the experiment. I gratefully acknowledge generous financial support by the Carlsberg Foundation. All mistakes are mine.

Introduction

Standard economic theory predicts that agents care only about their own material payoff. However, even in settings that focus purely on outcomes – and not on intentions – experimental economist have found ample evidence in contradiction with this prediction. While some agents do act in a selfish manner, others apparently choose to sacrifice own payoff in order to e.g. decrease inequality or increase efficiency. Hence, it seems that agents' preferences are formed over not only on their own outcome but on the distribution of outcomes. This finding is important as theoretical models that aim to explain and predict behavior thus need to incorporate these finding if we are to trust the predictions of such models. As an example, Fehr and Fischbacher (2002) show that without accounting for distributional preferences it is not possible to understand effects of competition on market outcomes.

While there are several studies on distributional preferences (sometimes labeled social preferences because agents care not only about themselves but instead take a social perspective), there is no golden rule as to how to define or measure these preferences. A widely used approach is to focus on a few forms of distributional preferences; to specify a parametric model and to use a variety of dictator and ultimatum games to estimate the parameters of the model. Results, of course, depend on the types of preferences included and on the functional form of the model.

We use a non-parametric approach suggested by Kerschbamer (2010). By definition, we do not specify a behavioral, parametric model but instead we use a series of dictator games to elicit the slopes of subjects' indifference curves in the own/other-payoff space. Subjects' indifference curves allow us to infer the sign of the effect on their utility from changes in other subjects' payoffs (e.g. whether a subject's utility increases, decreases or is constant when a different subject's payoff increases). In return, these effects map into the complete set of nine possible archetypes of distributional preferences.

In particular, we apply Kerschbamer's (2010) XY test on a large sample of the Danish population to make three contributions to the literature of distributional preferences. The first contribution is methodological and investigates the effect of role certainty. In the XY test, decision makers choose between income distributions that determine both the decision maker's own income and the income of a passive recipient. We employ two treatments relating to role certainty. In treatment FixedRoles, roles are determined ex ante and

participants chosen to be decision makers know that their choices will affect the payoff of a recipient for sure (identical to the classic dictator game). Participants chosen to be recipients make no choices and cannot affect outcomes. The dictator game is chosen as method because the test focuses purely on outcomes and not on intentions. In treatment RandomRoles, all participants make choices as if they are decision makers and actual roles are randomly determined ex post. Hence, the choices of half of the decision makers are inconsequential ex post but not ex ante. This procedure allows us to elicit the distributional preferences for all participants but the test is less clear. For instance, one could imagine that intentions matter in the sense that participants take beliefs about others' intended behavior into account when making their own choices.

Second, this is the first time that the XY test has been used on a large sample and the first time that distributional preferences have been widely measured in Denmark. We perform the test over the internet which has several advantages. Most importantly, it ensures that we get participants from all walks of life. This might be important for the external validity of the test as the preferences of the standard student population might differ from those of the general population. We ask subjects to answer questions about their socio-economic backgrounds and attitude questions from the World Values Survey. They also complete tests of IQ, cognitive reflection and personality. The large, heterogeneous sample ensures that there is variation in the responses. Thus, our procedure enables us to study the correlation between distributional preferences and personal characteristics (socio-economic, attitudes and psychological).

Third, we investigate how distributional preferences affect cooperation. In addition to the distributional preferences test, subjects participate in a standard one-shot public good game, often used to study cooperation. We test the effect of pure distributional preferences on corporation in this setting where intentions and other group members' actions might also influence behavior.

The experiment is run on the internet Laboratory for Experimental Economics (iLEE) at the University of Copenhagen, Denmark. We collaborate with the official statistics agency (Statistics Denmark) and send out letters, inviting people to participate in an economic experiment over the internet. Participants are randomly selected from the Danish population. In total, 1,067 people log in on our homepage and complete the experiment between July and September, 2010. They log in using an ID-number generated by Statistics Denmark and

remain anonymous to us throughout the experiment. Participants are paid by bank transfer once the experiment is over.

The experiment yields several important findings. First, we find that the two treatments FixedRoles and RandomRoles yield results that are insignificantly distinguishable from each other. Hence, the RandomRoles design can be used to elicit the distributional preferences of all participants without biasing the result. This finding has implications for the implementation of the test and means that the XY test can be applied in standard lab settings to control for differences in distributional preferences across subjects and subject pools.

Second, we find that, in line with previous literature, the most common preference type (32 percent) is efficiency maximizers who are willing to give up own income to increase the income of the recipient. 23 percent of subjects act in a way that is consistent with inequality aversion, 20 percent with selfishness and 14 percent are classified as having maximin preferences. In total, these four most prevalent types make up more than 89 percent of decision makers. Thus, while the XY test encompasses a very comprehensive framework which allows for a full distinction between nine different preferences types, only four of the types seem to be important empirically. Relating preferences to personal characteristics, we find that women are less likely to be selfish and more likely to be inequality averse. A long tertiary education and an attitude that individuals and not the state are responsible for their own happiness are (weakly) correlated with the probability of being classified as selfish. A positive attitude for competition and a belief that people in general will treat one fairly is positively correlated with efficiency maximizing behavior. Higher scores in the IQ and CR tests are negatively related to being inequality averse while higher CR scores are positively correlated to being efficiency maximizing. In summary, we find that personal characteristics do correlate with distributional preferences and that personal characteristics can predict preference types fairly well.

Third, we find that efficiency maximizers and decision makers with inequality aversion or maximin preferences contribute more to the public good than those who are selfish. In addition, the public good game reveals a substantial discrepancy between the prediction of standard economic theory and observed behavior. We find that taking distributional preferences into account can explain 43 percent of the difference between standard economic theory and observed behavior. These findings can be seen as a validation of the XY test as they demonstrate the test's ability to predict behavior in other settings.

The paper is organized as follows. Section 2 discusses related literature, section 3 provides a short introduction to the XY test, section 4 explains the experimental design in detail and section 5 presents the results. Section 6 concludes and discusses our findings.

2 Measuring distributional preferences

Most theories of distributional preferences aim to model a particular type of preferences such as inequality aversion (e.g. Fehr and Schmidt 1999, Bolton and Ockenfels 2000) or altruism (e.g. Andreoni and Miller 2002). These theories normally modify the utility function of an agent such that it includes not only the agent's own payoff but also some element of the payoff distribution (e.g. the mean payoff of all agents)[1]. Sometimes, experimental techniques are then used to estimate the parameters in the utility function by letting agents take part in, for instance, dictator and ultimatum games. The estimated parameters translate into a particular type of distributional preference. In addition to including outcome distributions, some models also include elements related to intentions (e.g. Charness and Rabin, 2002, who adds an element of reciprocity to an inequality aversion model). In the following, we highlight a few examples of experimental results on outcome based models.

Charness and Rabin (2002) is a great example of eliciting distributional preferences from subjects' behavior in the lab. They ask subjects to play up to 32 variations of dictator and trust games. They consider three main types of distributional preferences: competitiveness, inequality aversion and efficiency maximization (and selfishness which is embedded in all three). They formulate a decision-making model which allows for altruism by assigning weights to the payoff of the other person (which can depend on whether the decision maker is ahead or behind, payoff wise) and a term that captures reciprocity. In the dictator games where reciprocity plays no role, they find that 97 percent of observations are consistent with efficiency maximizing behavior. 75 percent are consistent with inequality aversion, 68 percent with selfishness and 60 percent with competitive preferences. Hence, they are not fully able to distinguish between the different types. When looking at the subset of games where self-interest plays no role, they find that 70 percent are efficiency maximizing, 20 percent are inequality averse and 10 percent are competitive. Their overall conclusion is that efficiency maximization is a greater driver for decisions than inequality aversion.

[1] See Kerschbamer (2010) for a review of the theoretical literature on distributional preferences.

Andreoni and Miller (2002) use variations of dictator games to investigate whether subjects who might seem irrational because of non-selfish behavior act in a way that is in fact in accordance with Varian's (1982) Generalized Axiom of Revealed Preference (GARP)[2]. First, they find that less than 2 percent of subjects violate GARP. The 98 percent rational subjects differ widely in behavior as preferences are heterogeneous. In particular, 23 percent behaves perfectly selfishly, 14 percent have perfect maximin preferences and 6 percent act as if own and others' payoffs were perfect substitutes (there is a price to allocating payoff to either subject. In the case of perfect substitutes, the entire endowment is allocated to the subject with the lowest price). In order to classify the remaining participants, they come up with a weaker definition of the three preference types. In this classification, 47 percent are selfish, 30 percent have maximin preferences and 22 percent act in accordance with payoffs being perfect substitutes. In addition, they estimate CES utility function parameters for the three types and use this to predict behavior in out-of-sample games, including public good games. They find some accordance between their prediction and the first round and average contribution of repeated public goods games.

Fishman, Kariv and Markovits (2007) extend the experiment of Andreoni and Miller by letting subjects make decisions in not just 2-player but also 3-player dictator games. Again, they find a large degree of heterogeneity among subjects. In the 2-player setting, they find that 15 percent act in accordance with selfish preferences. Of the 45 subjects with rational, non-selfish preferences, 5 percent have perfect substitutes preferences, 11 percent have Cobb-Douglas preferences and 5 percent have maximin preferences. 49 percent show a preference for efficiency and 31 percent have a preference for minimizing inequality. As Charness and Rabin (2002), they find that concerns about efficiency seem to be more important than concerns about inequality and that this holds both when subjects make decisions involving their own payoff (2-player game) and when decisions affect only the payoffs of other subjects (in the 3-player game).

Engelmann and Strobel (2004) use three-person dictator games to compare the relative importance of efficiency maximization, maximin preferences, and inequality aversion. They find that inequality aversion does not explain behavior well while a combination of efficiency

---

[2] If A and B are distinct bundles and A is indirectly revealed preferred to B then B cannot be strictly directly revealed preferred to A if behavior is in line with GARP.

maximization, maximin preferences and selfishness can rationalize most of the observed behavior.

Engelmann and Strobel (2007) report results from an internet experiment with 1,103 participants. Most participants take part in ten three-person dictator games. Engelmann and Strobel investigate in how many of the ten games, a particular subject's behavior is consistent with seven different types of distributional preferences (competitiveness, efficiency maximization, envy, generosity, inequality aversion, maximin preferences and selfishness). Only three types of preferences have subjects that are consistent in all ten games: maximin (15 percent), efficiency maximizers (8 percent) and selfishness (10 percent). The rest of the participants behave in a way that is consistent with different preference types in different of the 10 games. Again, efficiency concerns seem to be more relevant than inequality aversion.

In summary, there is no golden standard as to which types of distributional preferences to test for, how to define the different kinds of preferences (for instance, subjects that equalize payoffs might be inequality averse in one study but may be maximizing efficiency in another) or how to test for distributional preferences. As such, it is difficult to compare outcomes across studies. Obviously, a test that classifies subjects as one of four types yields different results than a test that classifies subjects as one of nine different types[3]. However, one general result is that efficiency maximization is an important determinant of behavior.

3 The XY test of distributional preferences

This section provides a short introduction to the XY-test proposed by Kerschbamer (2010). We briefly summarize the assumptions underlying the test, the nine archetypical preference types and the methodology. For further details, the reader is referred to Kerschbamer (2010).

Kerschbamer considers a two person setting where participants make choices that determine their own monetary payoff (denoted m for "my") and the payoff of a randomly selected other participant, the recipient (denoted o for "other"). In the XY test, subjects make a series of dictator choices between income distributions (m, o) with the aim of eliciting the slope of an indifference curve in the (m, o) space. The slope of the indifference curve then translates into a preference classification with nine different archetypical preference types.

[3] This is either because all subjects are forced into fewer categories or because more subjects are classified as "other" and then discarded which changes the relative share of the included categories.

There are three assumptions underlying the test and the preference classification resulting from it. The first assumption is *ordering*, i.e. that preference relations on income allocations (m, o) can be represented by a continuous utility function $u(m, o)$ for all possible income allocations $(m, o) \in \Re^2$. This assumption ensures that preferences are complete and that alternatives can be ordered continuously. The second assumption is *strict m-monotonicity*: for given o, the utility of the decision maker is strictly increasing in m (i.e. $\partial u/\partial m > 0$ for all $(m, o) \in \Re^2$). Hence, the positive effect of an increase in the decision maker's own income dominates any negative affect that might arise from, for instance, increased inequality. Thus, decision makers should not be willing to simply burn money in order to decrease inequality. The third assumption is *piecewise o-monotonicity*: the general effect on the decision maker's utility (increase, decrease or indifferent) from a change in the income of the recipient depends only on whether the recipient has a higher or lower income than the decision maker and *not* on the size of the income difference (i.e. $\text{sign}(\partial u/\partial o)$ depends only on $\text{sign}(m - o)$ for all $(m, o) \in \Re^2$). Thus, only outcomes – and not intentions – matter for the utility of the participant. For instance, belief-based concerns (such as reciprocity or guilt aversion) or the context of the situation (e.g. entitlement) cannot affect utility under this assumption. Hence, for the test to create reliable outcomes the experiment should to be designed in a way that makes these aspects less likely to influence behavior.

Under the three assumptions above, distributional preferences can be classified into nine different archetypes: selfish, altruistic, inequality averse, maximin, equality averse, competitive, envious, kick-down and kiss-up[4]. Consider first the selfish type who cares only about own material payoff. The utility of this person is unaffected by changes in the material payoff of the other person ($\partial u/\partial o = 0 \; \forall \; u$). Hence, indifference curves in the (m, o) space are vertical. A person that is efficiency maximizing (or altruistic) is better off when the payoff of the other person is increasing, that is $u(m, o)$ is increasing in all o. This person's indifference curves are negatively sloped. Conversely, a competitive or spiteful person is better off when the other person decreases and thus has positively sloped indifference curves.

Inequality (or inequity) aversion is characterized by disutility from differences in material payoff. Thus, an inequality averse person has negatively sloped indifference curves ($\partial u/\partial o > 0$) in the domain of advantageous inequality (i.e. when the person's own payoff is larger than the payoff of the other person, $m > o$) and positively sloped indifference curves

[4] See Kerschbamer (2010) for a survey of papers that discuss the various forms of preferences.

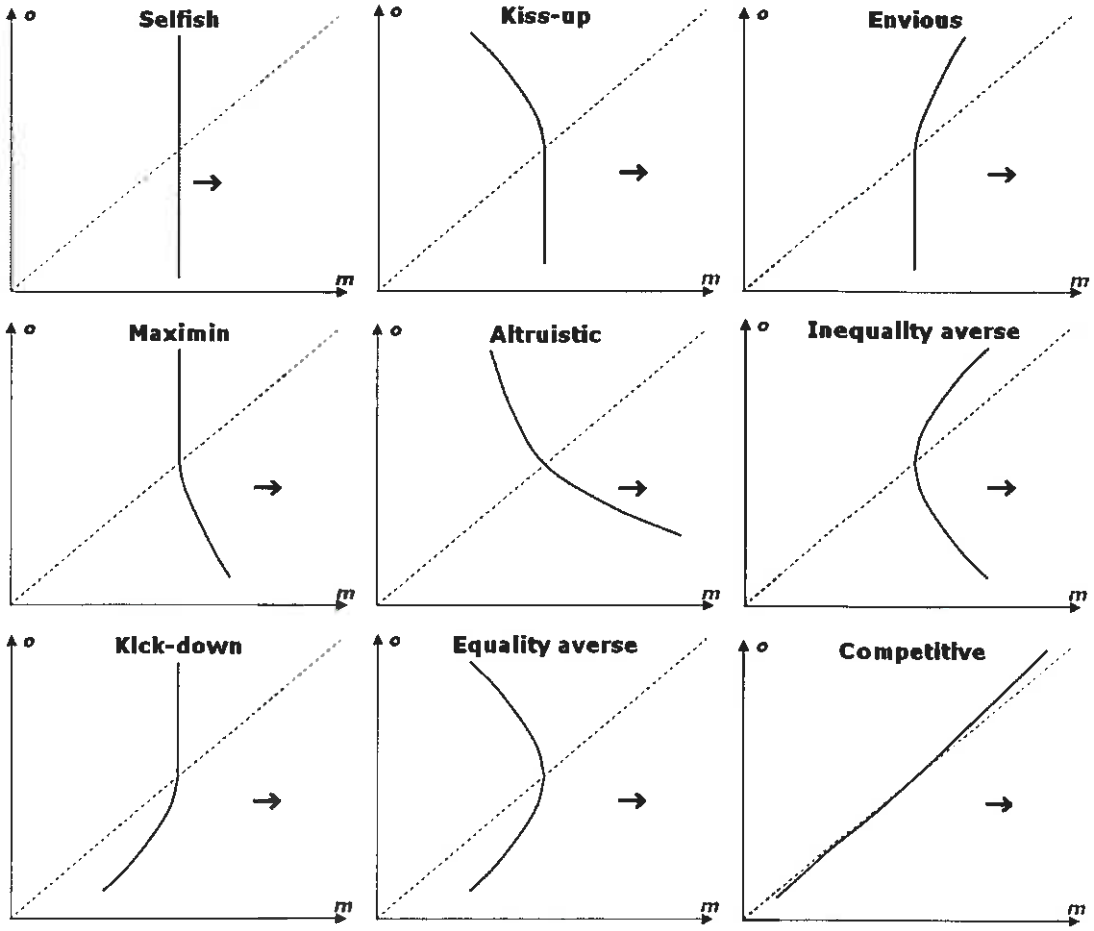
($\partial u/\partial o < 0$) in the domain of disadvantageous inequality ($m < o$)[5]. Conversely, a person that is equality averse has positively (negatively) sloped indifference curves in the domain of advantageous (disadvantageous) inequality.

The defining characteristic of maximin (or Leontief or Rawlsian) preferences is that the utility is increasing in the lowest payoff. That is, indifference curves are negatively sloped in the domain of advantageous inequality and vertically sloped in the domain of disadvantageous inequality. Conversely, a person with 'kiss-up' preferences has utility that is increasing in the highest payoff. An envious person is unaffected by the income of the other person if this is lower than the person's own income but gets disutility if the other person has more. Finally, the opposite is true for a person with 'kick-down' preferences who is unaffected if the other person has more but who gets increased utility from decreased payoff to the other person once this person has less. Figure 1 shows a graphical illustration of the nine archetypes.

To elicit the slope of the indifference curve – and thus to identify the distributional preferences – Kerschbamer lets participants make a series of choices between pairs of income distributions (m, o). Participants choose between Left and Right and in all situations, the option Right yields the egalitarian outcome in the reference point (m, o) = (e, e). The option Left is from one of two lists: the X-list (of disadvantageous inequalities) or the Y-list (of advantageous inequalities). In the X-list, the recipient is better off than in the reference point, (m, o) = (m', $e + g$), and in the Y-list the recipient is worse off, (m, o) = (m'', $e - g$).

[5] Inequality aversion is related to the strict egalitarian fairness ideal (see Cappelen et al. 2007) according to which total income should be distributed equally among all agents.

Figure 1: Indifference curves for the nine archetypes of distributional preferences



Notes: The figure shows the nine archetypes of distributional preferences. The arrows indicate the locus of the upper contour set (i.e. the allocations that make the decision maker better off than the points on the indifference curves).

In essence, the choices are dictator games as they determine both the decision maker's own payoff as well as the payoff of the recipient who has no influence on outcomes. The recipient is made passive in order to focus on distributional preferences and to eliminate the effect of intentions. Under the three assumptions, the choices allow us to elicit the slope of the participants' indifference curve through the reference point. This slope translates into one of the nine preference types, cf. figure 1.

4 Experimental design

This section first describes the detailed design of the distributional preferences test and our experimental treatments. Second, we briefly explain the overall procedures of the iLEE

internet experiments. Third, we describe two other parts of the experiment which we use in the analysis below: attitude questions and psychology measures. Finally, we give a brief summary of the design of the public good experiment which we use to measure the correlation between distributional preferences and cooperation and to test the predictive power of the distributional preferences test.

The distributional preferences test

We apply a modified version of the XY test of Kerschbamer (2010) described above. We set the reference point, e, to 50 Danish kroner (Dkr.) for each person (approximately €7) and table 1 shows the 14 distributions that decision makers can choose instead of the reference point. We make two modifications to the basic version of the XY test. First, we add two choices to the low end of the X-list and two to the high end of the Y-list, thus making the test asymmetric. Second, we vary the incremental change in m in the Left option (the step size in Kerschbamer's terminology, which is constant in the basic version of the test).

Table 1 shows the 14 pairs of distributions that decision makers choose between (7 in the X-list and 7 in the Y-list). Note that the assumption of *strict m-monotonicity* implies that decision makers should switch at most once from Right to Left and never from Left to Right.

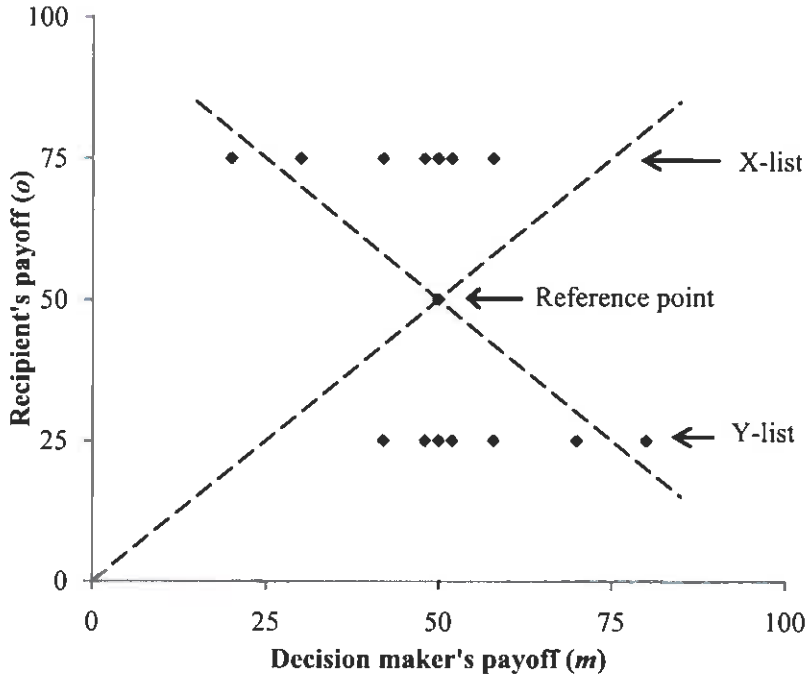
Table 1: The X- and Y-list used in the experiment

The X-list: Disadvantageous inequalities				**The Y-list: Advantageous inequalities**			
Left		Right		Left		Right	
m	*o*	*m*	*o*	*m*	*o*	*m*	*o*
20	75	50	50	42	25	50	50
30	75	50	50	48	25	50	50
42	75	50	50	50	25	50	50
48	75	50	50	52	25	50	50
50	75	50	50	58	25	50	50
52	75	50	50	70	25	50	50
58	75	50	50	80	25	50	50

Notes: The table shows the 14 set of income allocations (Left) that decision makers can choose instead of the reference point (Right). In each situation, they choose between Left and Right, where Right is always the reference point of 50 kr. for both the decision maker (denoted *m* for "my") and the unknown recipient (denoted *o* for "other"). For choices on the X-list, the decision maker always receive less than the participant if choosing Left (hence "disadvantageous inequalities). Analogously for the Y-list where the decision maker always receive more than the recipient ("advantageous inequalities").

The reason for using 50 kroner as the reference point is that this amount is equal to the endowment given to participants in the public good game which we use to test the predictive power of the test (see section 5.3 below). As noted by Levitt and List (2007), it is important to control for the stakes across games if we are to trust the predictions. We use an asymmetric version of the test to get allocations that are below the minus 45 degree line on the X-list and choices that are above the minus 45 degree line on the Y-list. This enables us to identify participants with strong altruistic concerns. These subjects are willing to give up more than the recipient gains (in at least one list) and thereby lowering efficiency. We vary the step size in order to have small steps close to the reference point while the steps increase in size away from the reference point. The small step size increases the precision with which we identify indifference curves that are (locally) vertically sloped. As the test allows us to identify two points between which the indifference curve must pass, we will never be able to find evidence of indifference curves that have exact vertical slopes. Reducing the step size gives us more precision in identifying those indifference curves that have slopes that are close to vertical. In principle, we define what we interpret as vertical. We set the step size equal to 2 Dkr. at m = 50 which allows us to identify slopes that in absolute terms are greater than 12.5 (25/2) and we consider these as being vertical. Hence, our definition implies that decision makers that are not willing (in either list) to give up 2 Dkr. in order to increase the payoff of the recipient by 25 Dkr. are classified as being selfish (their absolute willingness to pay to increase o by 1 is less than 1/12.5 = 0.08). Consider, for example, a decision maker that switches from Right to Left at m = 52 on both lists. This behavior is in principle consistent with both inequality aversion and selfishness. However, we say that the inequality aversion is so weak (2 Dkr. in monetary gains to the decision maker is enough to compensate for an inequality of at least 23 Dkr.) that we label this behavior as purely selfish. Increasing the step size away from m = 50 ensures that we have distributions with m-values that are further from 50 without having to increase the number of choices by each decision maker. Figure 2 provides an illustration of the income allocations in table 1.

Figure 2: Possible income distributions



Notes: The figure shows the 14 income distributions that decision makers can choose instead of the reference point (which gives 50 kr. to both the decision maker, m, and the recipient, o).

Experimental schedule

Participants are first explained the rules of the experiment (see appendix A for the instructions). Decision makers then choose between the 14 pairs of distributions in table 1 but only one is paid (which one is determined randomly ex post). Choices are made one at the time on separate screens where decision makers choose between Left and Right before moving on to the next choice (see appendix B for screen shots). The order of the choices is randomized. Once they have made all 14 choices, they see a confirmation screen. The confirmation screen shows an overview of all decisions with a horizontal line separating the X- and the Y-list. The chosen distributions are color highlighted and decision makers can go back and change their decisions as many times as they wish. Once they confirm their decisions, they move on to the next experiment in the wave[6].

[6] The iLEE experiments are carried out in waves. Our distributional preferences test is part of the third wave which in total consists of six different parts.

Treatments

We employ two primary treatments that relate to the roles and possible interaction of decision makers and recipients. The treatment FixedRoles is exactly as described in section 3 where half of the participants are decision makers, the other half are recipients and the two are randomly matched. Roles are randomly assigned and revealed after both participants have read the instructions. Recipients make no decisions and continue on to the next part in the wave. In treatment RandomRoles, all participants make choices as if they are decision makers. A random draw ex post determines which role each participant will be paid as. Subjects chosen to be decision makers will then be randomly matched with those chosen to be recipients. Instructions are kept as similar as possible across treatments. Treatment allocation is random with 1/3 of participants in treatment FixedRoles and 2/3 of participants in RandomRoles.

In addition, we run four secondary treatments in a 2x2 design relating to the presentation of the X- and Y-lists on the confirmation screen (only shown to subjects who make decisions). The first dimension is whether the X- or the Y-list is shown first and the second dimension is the ordering (ascending or descending) within lists. Allocation to treatments is random and treatments are equally likely.

Treatment FixedRoles is the clearest way to isolate the effect of distributional preferences on behavior and it is the best way to elicit these preferences. The reason is that the decision maker knows – when making the decisions – that the recipient is a passive person who cannot influence the payoffs of neither the decision maker or a third participant. In addition, the fact that the recipient's role is fixed from the outset means that concerns about intentions (such as reciprocity) cannot play a role. Hence, the decision maker can base decisions purely on the basis of distributional preferences over outcomes. The main downside to treatment FixedRoles is that we elicit preferences for only half of the participants. This means that it is costly to collect a large sample. In addition, if the test is to be used in laboratory settings to control for heterogeneity in distributional preferences it would be preferable if the preferences could be elicited for all participants. This is possible in the RandomRoles treatment where all participants make decisions as decision makers and roles are randomly assigned ex post. In this case, however, one could imagine that beliefs about other's actions influence decisions. Also, it could be a concern that participants do not think as much about their choices given that they are not certain to be influential. One aim of this paper is to see whether these two procedures yield different results. Finally, we vary the

ordering and sequence of the X- and Y-lists to control for any behavioral patterns resulting from presentation.

General iLEE procedures

The experiment is conducted using the platform of the internet laboratory for experimental economics (iLEE) at the University of Copenhagen, Denmark. Subjects for the platform are recruited with the assistance of the official statistics agency (Statistics Denmark) who select a random sample from the general population. Statistics Denmark send the selected individuals physical letters, inviting them to participate in an online scientific experiment that is jointly organized by the University and Statistics Denmark. Participants log in to the experiment using a personal identification code provided by Statistics Denmark. Payments are done by electronic bank transfer and participants remain anonymous to the researchers at the University throughout the experiment. The distributional preferences test is part of the third wave of experiments conducted on the iLEE platform. All three waves are done on the same set of participants, thus creating a panel data set useful for cross game analysis. We use this feature to examine the predictive power of the measure of distributional preferences. For the third wave, we invite the 2291 people who completed the first wave[7]. In total, 1067 participants complete the third wave between July and September, 2010. Participants can log on at any point during this period and are free to log out and continue later at their convenience. The third wave consists of a total of 6 different parts. First in the third wave is a trust game. After the trust game are four different smaller parts: a real effort task, a voting game, measures of risk and loss aversion and our application of the XY test. The order of these four parts is random. The final part is a questionnaire which includes questions on age, gender and education. In total, the median person spends 63 minutes completing the entire wave and earns 279 Dkr. (€37).

Cooperation with Statistics Denmark is necessary to obtain the names and addresses of participants needed to send out invitations but our cooperation also yields additional advantages. First, it allows us to target a representative sample of the population. Combined with the high penetration of internet access in Denmark[8], this means that we have participants

---

[7] For descriptive statistics on the sample and the full details about invitations and general iLEE procedures, see Tyran and Wengström (2009) as well as the online appendix for that paper.

[8] In 2009, Denmark was the world leader in terms of broadband internet penetration (source: EU Commission's 15th Progress Report on the Single European Electronic Communications Market, 2009)

from all walks of life. This enables us to investigate how experimental behavior is correlated with self-reported socio-economic variables such as age, education and employment. Second, our procedures entail double blindness in the sense that participants are anonymous not only to other participants but also to us, the experimenters. This is important to minimize potential experimenter-demand effects. Levitt and List (2007) survey evidence that shows how the lack of anonymity between experimenters and participants increases the level of pro-social behavior when measuring distributional preferences. Double blindness should decrease such effects.

Questionnaire

Participants answer questions regarding their basic socio-economic background, including their age, gender and level of education. In the analysis below, we group education in four categories: primary (no more than 10 years, 6 percent), secondary (vocational and high school, 22 percent), short tertiary (50 percent) and long tertiary (22 percent). In addition, we ask participants to answer five attitude questions from the World Values Survey.[9] Participants had the option of not answering the questions. About 8 percent choose to not answer at least one of the following five questions:

LeftRight: "In political matters, people talk of "the left" and "the right." How would you place your views on this scale if 1 means the left and 10 means the right?" Possible answers are integers ranging from 1: "left" to 10: "right".

Responsibility: "We would like your opinion on important political issues. How would you place your views on a scale from 1 to 10?" Possible answers are integers ranging from 1: "People should take more responsibility to provide for themselves" to 10: "The government should take more responsibility to ensure that everyone is provided for".

Competition: How would you place your views on a scale from 1 to 10?" Possible answers are integers ranging from 1: "Competition is good. It stimulates people to work hard and develop new ideas" to 10: "Competition is harmful. It brings out the worst in people".

Trust: "Generally speaking, would you say that most people can be trusted or that you can't be too careful in dealing with people?" Possible answers are 0: "Can't be too careful" and 1: "Most people can be trusted".

---

[9] See http://www.worldvaluessurvey.org/.

Fairness: "Do you think most people would try to take advantage of you if they got a chance, or would they try to be fair?" Possible answers are integers ranging from 1: "Would take advantage of you" to 10: "Would try to be fair".

Psychology measures

We also include psychology measures in the survey that participants answer. Our psychology measures consist of a cognitive reflection test (CRT), an IQ test and a personality test. The CRT is due to Frederick (2005) and consists of three short questions that all have incorrect but "intuitive" answers. Hence, the CRT is aimed at capturing participants' ability to reflect upon a question and resist the temptation of giving the first (wrong) answer that comes to mind. Frederick finds that the CRT is predictive of behavior in a number of decision making environments. The three questions are:

1: "A bat and a ball cost 110 Dkr. in total. The bat costs 100 Dkr. more than the ball. How much does the ball cost?" Answer is given in Dkr.

2: "If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?" Answer is given in number of minutes

3: "In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake?" Answer is given in number of days.

The intuitive answers to the questions are (10, 100 and 24) while the correct answers are (5, 5 and 47). The variable *CRT score* is calculated as the number of correct answers, i.e. 0, 1, 2 or 3.

Our measure of IQ is based on the I-S-T 2000R intelligence structure test (which we use by permission of Dansk Psykologisk Forlag[10]). The test is based on Raven's Progressive Matrices and participants have 10 minutes to solve 20 puzzles (see appendix B for a screen shot of an example). The *IQ score* variable is the number of correct answers, from 0 to 20.

Our measure of personality is based on the Five Factor Model (Costa and McCrae 2004) which describes human personality according to the "Big Five" dimensions or traits: *Openness* (to experience), *Conscientiousness, Extraversion, Agreeableness* and *Neuroticism*. *Openness* is related to creativity, to being curios and original and to the person's ability to

[10] See http://www.dpf.dk/.

contemplate new ideas. *Conscientiousness* is related to having a will to achieve, to being conscientious, hard-working and well-organized and to being ambitious. *Extraversion* is related to being social, passionate, talkative and dominating in groups. *Agreeableness* is related to kindness and altruism and to being good-natured and trusting. *Neuroticism* is related to being emotional, worried, self-conscious and temperamental. We use the short version of the NEO PI-R test[11] with 60 questions. The test yields scores for each of the five dimensions on a scale from 1 to 48. A higher score means that a personality is correlated with a higher degree of the particular trait. For example, a person who scores 40 on *Neuroticism* is likely to be more emotional than a person who scores 5.

Public good experiment

In the public good experiment, subjects are endowed with Dkr. 50 and decide how much to contribute to the public good and how much to keep for themselves (the private good). Subjects are matched in groups of four. The total amount contributed by the group to the public good is doubled and shared equally among the group members (marginal per capita return, MPCR, is 0.5). It is a one-shot game as subjects make only one contribution decision. While it is socially optimal that all group members contribute the full endowment, individual income is maximized by contributing zero as there is no scope for reputation building. After the contribution decision, we elicit beliefs about the average contribution of the three other group members (incentivized using a quadratic scoring rule). In the analysis below, the variable *Contribution* is the contribution decision and *Belief* is the elicited belief.

5 Results

In this section, we present the results from the experiment. First, section 5.1 provides simple descriptive statistics and shows that the different treatments yield similar results. Second, section 5.2 investigates how the distributional preferences correlate with socio-economic background variables and personal traits like IQ, personality and gender. We find that gender, age, subjects' expectation of being treated fairly, cognitive reflection and IQ correlate with distributional preferences. Finally, section 5.3 tests the predictive power of the XY-test by comparing actual and predicted behavior in a standard public good game. We find that subjects who are efficiency maximizers, inequality averse and have maximin preferences

[11] The Danish version of the test is developed by Dansk Psykologisk Forlag, see Costa and McCrae (2004).

contribute more than those who are selfish. In addition, we find that taking distributional preferences into account explains almost half of the difference between observed behavior in the public good game and the prediction of standard economic theory.

5.1 The distribution of distributional preferences

In total, n = 1067 participants complete the distributional preferences test (average earning from this part is 51.8 Dkr.). Table 2 shows that 363 participate in the FixedRoles treatment, of which 181 are decision makers and 182 are recipients[12]. The 704 participants in the RandomRoles treatment all make decisions as if they are decision makers and a random draw determines the role and matching of participants ex post. We say that decision makers are rational if their behavior is consistent with the two assumptions of *ordering* and *strict m-monotonicity*. These assumptions imply that decision makers should switch at most once from Right to Left (and never from Left to Right) in either list. Of the n = 885 decision makers, 650 fulfill the rationality criterion while 235 (27%) make choices that are not consistent with rational behavior[13].

Table 2: Number of participants

Treatment	RandomRoles	FixedRoles		
Role	-	Decision makers	Recipients	*Total*
All participants	704	181	182	*1,067*
- Inconsistent	190	45	-	*235*
- Consistent	514	136	-	*650*

Notes: The table shows the number of participants split by treatments (RandomRoles and FixedRoles), role (decision maker and recipient in FixedRoles) and consistency (inconsistent and consistent for decision makers).

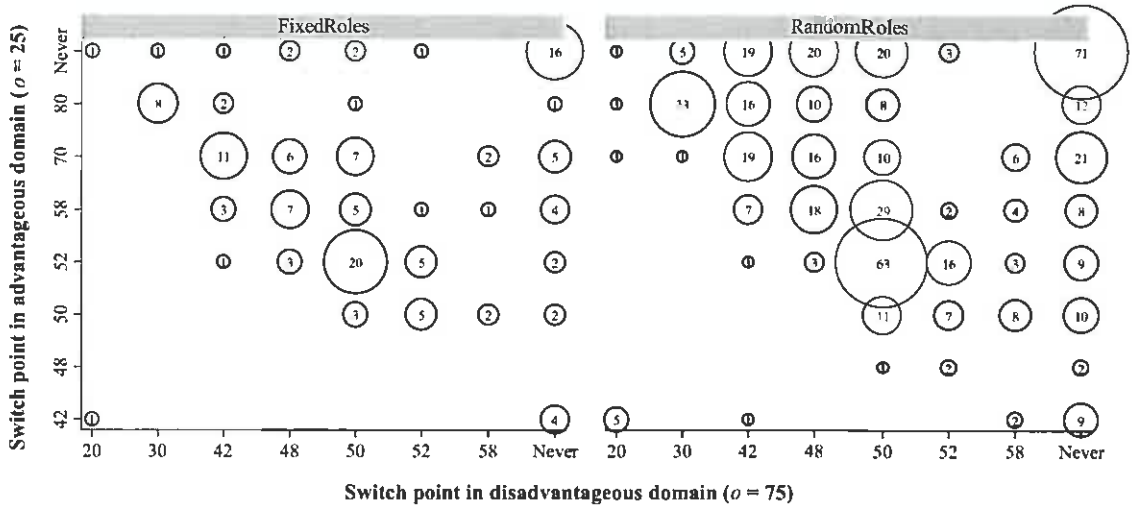
Figure 3 shows at what level of own income (m) decision makers switch from Right to Left in each of the treatments FixedRoles (n = 136) and RandomRoles (n = 514). The first thing to note from the figure is that the two treatments seem to yield similar results. In total, there are 8 x 8 = 64 possible combinations of switch points and of these, we observe 43. The distributions on the 43 observed switch points are not different for the two treatments (p = .592, χ^2 test). In addition, the distributions do not differ across the two other treatments

[12] The spare recipient is matched with a random decision maker whose choice is consequential twice.

[13] 272 subjects had inconsistent choices when they reached the confirmation stage. Of these, 40 changed choices in a consistent way while 3 subjects changed from consistent to inconsistent behavior.

relating to presentation on the confirmation screen: the order in which the table are presented ($p = .682$, χ^2 test) and the ordering within tables ($p = .566$, χ^2 test).

Figure 3: Switch points by treatment



Notes: The figure shows at what level of own income (m) decision makers switch from Right to Left, for the two treatments FixedRoles ($n = 136$) and RandomRoles ($n = 514$). The x-axis shows the switch point for the X-list where the recipient receives $o = 75$ (disadvantageous domain) and the y-axis shows the switch point for the Y-list where the recipient receives $o = 25$ (advantageous domain). The reference point of Right is the egalitarian distribution of (50, 50). Marker size is proportional to the number of observations at a particular point (also printed). For the disadvantageous domain, switching at $m \leq 48$ implies a negatively sloped indifference curve above the reference point ($o > 50$), switching at $m = \{50, 52\}$ implies a vertical sloped indifference curve and switching at $m = 52$ or never switching implies a positively sloped indifference curve. Analogously for the advantageous domain, switching at $m \leq 48$, implies a positively sloped indifference curve below the reference point ($o < 50$), switching at $m = \{50, 52\}$ implies a vertical sloped indifference curve and switching at $m > 52$ or never switching implies a negatively sloped indifference curve.

We use the switch points in figure 3 to classify subjects according to their distributional preferences and the result is presented in table 3. For example, a person switching at $m \geq 58$ on both lists behaves in a way that is consistent with inequality aversion. The second column of the table shows that about one third of the subjects are classified as efficiency lovers, about one quarter are inequality averse and one fifth are selfish. Almost 14 percent have maximin preferences and in total these four preference types account for 89 percent of the subjects. Of the remaining, 6 percent act in a way that is consistent with envy, 3 percent are spiteful while kiss-up, equality averse and kick-down each account for only about 1 percent of the subjects. Thus, while the XY is very comprehensive and allows for the full set of nine theoretical preferences types, only four of the types seem to be important empirically. The fact that the X- and Y-lists crosses the minus 45 degree line allows us to identify among those classified as

efficiency lovers the subjects that must have some element of pure altruistic preferences. In particular, subjects switching at $m = 20$ on the X-list give up more than the recipient gains (compared to the reference point) and thus efficiency is decreased. Similarly, never switching on the Y-list is inconsistent with efficiency concerns (in the last row, the decision maker chooses a total output of 100 instead of 105). This behavior is observed for 52 of the 209 subjects (25%) classified as efficiency lovers. Hence, for these subjects altruism must be part of their preferences as their choices are inconsistent with pure efficiency maximization.

Columns three and four show the distributions for the two treatments RandomRoles and FixedRoles. A Fisher's exact test fail to reject that the two distributions are the same ($p = .549$). The fifth column shows the expected distribution if all decision makers were to choose switch points randomly (and each switch point has the same probability of being chosen). The distribution resulting from random behavior is significantly different from the overall observed distribution ($p = .000$, χ^2 test). In addition, we test whether the two treatments relating to presentation makes a difference for the observed distribution of preference types. We find that neither the table order ($p = .679$, Fisher's exact test) nor the questions ordering ($p = .555$, Fisher's exact test) matters for the results. Hence, for the remainder of the analysis we merge all treatments.

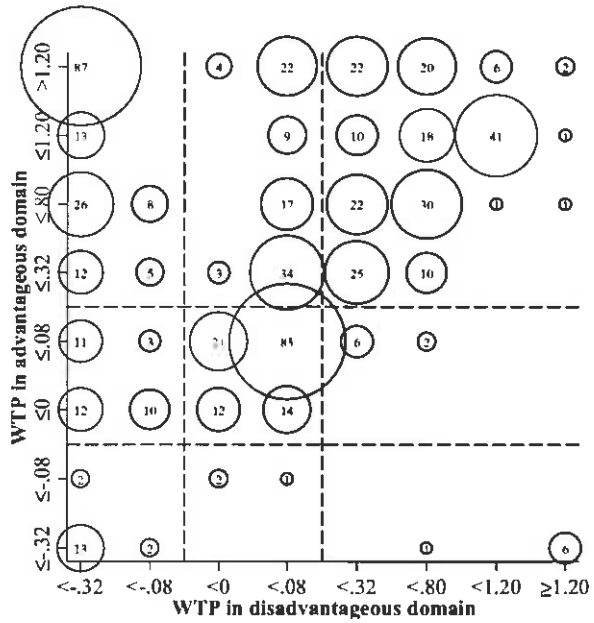
Table 3: Distribution of distributional preference types (%)

	Overall	RandomRoles	FixedRoles	Random switch
Efficiency loving	32.2	32.5	30.9	25.0
Inequality averse	23.2	23.7	21.3	12.5
Selfish	20.0	18.9	24.3	6.3
Maximin/Leontief	13.7	14.0	12.5	12.5
Envious	5.5	5.8	4.4	6.3
Spiteful	2.6	2.5	2.9	6.3
Kiss-up	1.2	0.8	2.9	12.5
Equality averse	1.1	1.2	0.7	12.5
Kick-down	0.5	0.6	0.0	6.3
N	650	514	136	

Notes: The table shows the distribution (in %) of distributional preferences for the total sample ($n = 650$) as well as split by the two treatments RandomRoles ($n = 514$) and FixedRoles ($n = 136$). A Fisher's exact test fails to reject the hypothesis that the distributions are the same for the two treatments ($p = .549$). The rightmost column shows the expected distribution if decision makers choose switch points randomly. This distribution is significantly different from the observed overall distribution ($p = .000$, χ^2 test).

In addition to classifying subjects as types, we also calculate their willingness to pay (WTP) in order to change the income of the recipient by 1. For example, subjects switching from Right to Left at m = 42 on the X-list are willing to give up at least 8 but not 20 to increase the income of the recipient by 25. Hence, these subjects have $0.32 \leq \text{WTP} < 0.80$. Similarly, subjects that switch at m = 58 on the X-list are willing to give up at least 2 but less than 8 in order to decrease the income of the recipient by 25. Thus, for these subjects $-0.32 \leq \text{WTP} < -0.08$. Figure 4 shows the distributions of WTP in our experiment. A higher WTP means that a decision maker is more benevolent towards the recipient. The figure also shows a grid to divide subjects into the nine distributional preferences types reported in table 3. For instance, efficiency maximizing agents have WTP > 0.08 is both domains and are located in the top-right corner of the figure (see the figure's notes for the location of the other types).

Figure 4: Willingness to pay to change the income of the recipient



Notes: The figures shows the decision makers' willingness to pay (WTP) for changing the income of the recipients (n = 650). The x-axis shows the WTP in the disadvantageous domain (where $o = 75 > m$) and the y-axis shows the WTP for the advantageous domain (where $o = 25 < m$). A *positive* WTP is the amount that the decision maker is willing to give up in order to *increase* the income of the recipient by 1. Analogously, a *negative* WTP is what the decision maker is willing to give up to *decrease* the income of the recipient by 1. We calculate intervals for the WTP and the figure shows the upper bounds (except for when WTP ≥ 1.20). For example, the label <-0.08 means $-0.32 \leq \text{WTP} < -0.08$. The marker size is proportional to the number of observations at a particular point (also printed). The dotted lines split the figures in three columns and three rows. Label the columns A, B and C and the rows 1, 2 and 3. Then A1 contains the inequality averse, A2 the maximin, A3 the efficiency lovers, B1 the envious, B2 the selfish, B3 the kiss-up, C1 the spiteful, C2 the kick-down and C3 the equality averse.

The figure shows that, in the advantageous domain, few subjects are willing to give up money in order to decrease the income of the recipient (27 of 650, about 4 percent, have a negative WTP). In addition, it shows that in the advantageous domain (when o = 25) 25 percent (163 of 650 participants) have a WTP of at least 1.20. Hence, they are willing to give up more of their own income than the recipient gets and thus lower overall output. These people are either very inequality averse, very altruistic or have strong maximin preferences.

Our findings are in line with the previous literature in the sense that we find greater support for a motive of efficiency maximization than for inequality aversion. Our finding of 20 percent selfish subjects is close to the results reported by Andreoni and Miller (2002) who find that 23 percent are selfish. While our findings are consistent with previous findings, one should be careful in putting too much weight on cross-study comparisons as definitions differ across papers.

5.2 The correlates of distributional preferences

We use the heterogeneity of our sample to investigate how the different preference types correlate with background variables for the four most prevalent types (selfish, inequality averse, maximin and efficiency maximizers, n = 579). In the first step, we include only *basic* socio-economic variables often collected in experiments (gender, age and dummies for education). Table 6 presents the result in form of average marginal effects based on a multinomial logistic regression where the dependent variable is the type classification (*Type*).

Table 4: Correlates of distributional preferences (basic)

Dependent variable: Type	Selfish	Inequality averse	Maximin	Efficiency maximizers
Female	-0.082**	0.141***	0.073**	-0.132***
	(0.035)	(0.034)	(0.029)	(0.039)
Age	0.003	0.020***	-0.010*	-0.013
	(0.007)	(0.008)	(0.006)	(0.008)
Age squared	-0.032	-0.152*	0.079	0.105
	(0.072)	(0.078)	(0.065)	(0.086)
Secondary education	0.153	-0.045	0.006	-0.114
	(0.097)	(0.079)	(0.075)	(0.097)
Short tertiary education	0.092	-0.095	0.022	-0.019
	(0.094)	(0.071)	(0.070)	(0.091)
Long tertiary education	0.177*	-0.143*	-0.034	0.001
	(0.095)	(0.078)	(0.076)	(0.095)
Estimated probability	22.5	26.1	15.4	36.1
Actual share	22.5	26.1	15.4	36.1
Percent correctly predicted	2.5	49.7	7.4	72.1
N	579			
Pseudo log-likelihood value	-738.2			
Pseudo-R2	0.050			

Notes: The table shows average marginal effects from a multinomial logistic regression with robust standard errors (shown in parentheses). The dependent variable is the *Type* classification from *table 3*. Basic independent variables include a gender dummy, the participants' *age* and the *age squared* (scaled up by a factor 1'000) and education dummies. Estimated probability is the predicted share of each type and actual share is the share found in the sample. Percent correctly predicted is the share of people that is predicted to be the same type as is observed. $N = 579$ as the regression only includes the four most prevalent types. * denotes significance at 10 percent, ** at 5 percent and *** at 1 percent.

The first row of table 4 shows that female subjects are less likely to be selfish (-8.2%) or efficiency maximizing (-13.2%) and more likely to be inequality averse (14.1%) or have maximin preferences (7.3%). The second row shows that age is positively related to being inequality averse (the effect is decreasing, third row of table 4) and negatively related to having maximin preferences. Education seems to have only a very weak effect on distributional preferences where a long tertiary education positively affects the probability of being selfish (17.7%) and negatively affects the probability of being inequality averse (-14.3%). The model performs very well in predicting the expected shares of each preference type but less well on the individual level (percent correctly predicted ranging from 2.5 to 72.1).

In addition to asking subjects basic socio-economic questions, we also ask them five attitude questions from the World Values Survey: *LeftRight* (identification as left or right wing on a scale from 1 to 10 where 1 is left and 10 is right), *Responsibility* (responsibility for

the individual is the person's own responsibility (1) or the governments (10)), *Competition* (attitude to competition from good (1) to bad (10)), *Trust* (can people generally be trusted, no (0) or yes(1)) and *Fairness* (will people treat you fairly, from 1 (no) to 10 (yes)). Table 5 shows the result of including these variables in the regression (n = 536 as 43 subjects did not answer at least one of these questions).

First, table 5 shows estimates for the basic control variables that are similar to those in table 4 (although some of the estimates are slightly smaller when attitudes are included). Second, the table shows that subjects who think that competition is good are more likely to be efficiency maximizers (*Competition* has a negative coefficient). Third, the last row of estimates show that subjects who expects others to treat them fairly are more likely to maximize efficiency and less likely to be selfish. Finally, the regression model with attitude questions yield better individual predictions of preference types, especially for the selfish subjects (20.5 percent are correctly predicted compared to 2.5 percent in table 4).

Table 5: Correlates of distributional preferences (basic and attitudes)

Dependent variable: Type	Selfish	Inequality averse	Maximin	Efficiency maximizers
Female	-0.084**	0.134***	0.059*	-0.109***
	(0.037)	(0.037)	(0.032)	(0.042)
Age	0.004	0.017**	-0.008	-0.013
	(7.011)	(7.889)	(6.196)	(8.081)
Age squared	-0.036	-0.120	0.050	0.106
	(0.000)	(0.000)	(0.000)	(0.000)
Secondary education	0.121	-0.043	-0.010	-0.068
	(0.096)	(0.081)	(0.075)	(0.099)
Short tertiary education	0.061	-0.085	0.009	0.014
	(0.093)	(0.074)	(0.070)	(0.092)
Long tertiary education	0.174*	-0.137*	-0.050	0.012
	(0.095)	(0.080)	(0.075)	(0.097)
LeftRight	0.009	-0.019*	0.009	0.001
	(0.009)	(0.011)	(0.008)	(0.012)
Responsibility	-0.016	0.003	0.008	0.005
	(0.010)	(0.011)	(0.009)	(0.011)
Competition	0.015	0.006	0.008	-0.029**
	(0.011)	(0.011)	(0.009)	(0.012)
Trust	-0.093	0.056	-0.046	0.083
	(0.067)	(0.082)	(0.060)	(0.101)
Fairness	-0.030**	-0.012	-0.009	0.051***
	(0.012)	(0.013)	(0.010)	(0.016)
Estimated probability	22.8	26.7	15.1	35.4
Actual share	22.5	26.1	15.4	36.1
Percent correctly predicted	20.5	49.0	7.4	73.2
N	536			
Pseudo log-likelihood value	-665.0			
Pseudo-R2	0.076			

Notes: The table shows average marginal effects from a multinomial logistic regression with robust standard errors (shown in parentheses). The dependent variable is the *Type* classification from *table 3*. Basic independent variables are a gender dummy, the participants' *age* and the *age squared* (scaled up by a factor 1'000) and education dummies. In addition, five attitude questions from the World Values Survey are included (*LeftRight*, *Responsibility*, *Competition*, *Trust* and *Fairness*). Estimated probability is the predicted share of each type and actual share is the share found in the sample. Percent correctly predicted is the share of people that is predicted to be the same type as is observed. *N* = 536 as the regression only includes the four most prevalent types and 43 participants choose not to answer all attitude questions. * denotes significance at 10 percent, ** at 5 percent and *** at 1 percent.

In addition to the *basic* and *attitude* variables, we also include variables related to psychology. These include the number of correct answers to the IQ test (from 0 to 20, *IQ score*), the number of correct answers to the cognitive reflection test (from 0 to 3, *CR score*) and the score in the five domains of the Big 5 personality test (from 0 to 48, *Agreeableness, Conscientiousness, Extraversion, Neuroticism* and *Openness*). The latter five estimates are not presented in table 6 to save space (all are insignificant).

The first line of table 6 shows that once we control for psychological factors, female subjects are no longer more or less likely to be classified as either efficiency maximizers or as having maximin preferences. Thus, being female *per* se does not influence these probabilities but gender is instead correlated with underlying psychological characteristics. The effects of age, *Fairness* and *Competition* do not change much compared to table 5 while a long tertiary education is no longer correlated with the probability of being classified as inequality averse. *Responsibility* is negatively correlated with being selfish, i.e. subjects who think that the government should take more responsibility to ensure that everyone is provided for are less likely to be classified as selfish. Finally, higher scores in the CR test are positively related to being efficiency maximizing and negatively related to being inequality averse and the latter holds also for higher scores on the IQ test.

In conclusion, we find substantial correlation between social preferences and background variables (socio-economic, attitudes and psychology). These effects are of substantial magnitudes and most have the intuitively expected signs. Note also that the explanatory power of the model increases when controlling for psychology and attitudes. This is reflected both in the success of predictions at the individual (all except for the efficiency maximizers increase substantially) and aggregate level (which are all fairly accurate) as well as in the increase of log-likelihood values and R^2 (the latter doubles comparing table 6 to table 4).

Table 6: Correlates of distributional preferences (basic, attitudes and psychology)

Dependent variable: Type	Selfish	Inequality averse	Maximin	Efficiency maximizers
Female	-0.092**	0.104***	0.042	-0.055
	(0.041)	(0.039)	(0.036)	(0.043)
Age	0.002	0.018**	-0.009	-0.012
	(7.022)	(7.880)	(6.173)	(7.893)
Age squared	-0.025	-0.141*	0.062	0.104
	(0.000)	(0.000)	(0.000)	(0.000)
Secondary education	0.115	-0.012	-0.033	-0.070
	(0.096)	(0.083)	(0.077)	(0.096)
Short tertiary education	0.055	-0.052	-0.013	0.010
	(0.092)	(0.076)	(0.073)	(0.089)
Long tertiary education	0.171*	-0.075	-0.076	-0.020
	(0.094)	(0.081)	(0.077)	(0.094)
LeftRight	0.010	-0.017	0.008	0.000
	(0.010)	(0.011)	(0.008)	(0.012)
Responsibility	-0.018*	0.003	0.008	0.008
	(0.010)	(0.011)	(0.009)	(0.012)
Competition	0.016	-0.002	0.011	-0.025**
	(0.012)	(0.011)	(0.009)	(0.012)
Trust	-0.092	0.055	-0.043	0.080
	(0.069)	(0.079)	(0.062)	(0.101)
Fairness	-0.030**	-0.015	-0.010	0.055***
	(0.012)	(0.014)	(0.010)	(0.016)
IQ score	0.006	-0.014**	0.003	0.005
	(0.006)	(0.007)	(0.005)	(0.007)
CR score	-0.023	-0.041**	-0.016	0.080***
	(0.018)	(0.017)	(0.015)	(0.020)
Big 5 personality scores	Yes	Yes	Yes	Yes
Estimated probability	22.8	26.7	15.1	35.4
Actual share	22.5	26.1	15.4	36.1
Percent correctly predicted	26.2	52.4	12.3	67.4
N	536			
Pseudo log-likelihood value	-646.4			
Pseudo-R2	0.102			

Notes: The table shows average marginal effects from a multinomial logistic regression with robust standard errors (shown in parentheses). The dependent variable is the *Type* classification from *table 3*. Basic independent variables are a gender dummy, the participants' *age* and the *age squared* (scaled up by a factor 1'000) and education dummies. In addition, five attitude questions from the World Values Survey are included (*LeftRight*, *Responsibility*, *Competition*, *Trust* and *Fairness*) and measures related to psychology (scores in the IQ and cognitive reflection (CR) tests as well as Big 5 personality scores (not reported; all five are insignificant). Estimated probability is the predicted share of each type and actual share is the share found in the sample. Percent correctly predicted is the share of people that is predicted to be the same type as is observed. $N = 536$ as the regression only includes the four most prevalent types and 43 participants choose not to answer all attitude questions. * denotes significance at 10 percent, ** at 5 percent and *** at 1 percent.

5.3 Distributional preferences and cooperation

In this section, we investigate how distributional preferences can help explain behavior in the standard, one-shot public good game. We first briefly summarize the public good game before providing a descriptive overview of observed behavior. We then formalize the analysis using regressions. Finally, we predict behavior taking distributional preferences into account and compare these behavioral predictions to the predictions of standard economic theory (SET) and observed behavior. We find that a substantial share of subjects contribute positive amounts and that subjects classified as inequality averse, efficiency maximizers and those who have maximin preferences contribute more than those who are selfish, even when controlling for the beliefs about other subjects' contributions. Finally, we find that incorporating distributional preferences explains almost half of the difference between the prediction of SET and observed behavior. These findings can be interpreted as a validation of the XY test as it is able to predict behavior in a different setting.

In the public good game, participants in groups of four are endowed with 50 Dkr. and given the option to contribute to a common pot (the public good). The amount contributed to the pot is doubled and shared equally among all group members while the rest of the endowment is kept by subjects themselves (the private good). Hence, it is socially optimal for all group members to contribute their entire endowment but it is individually optimal to keep the endowment.

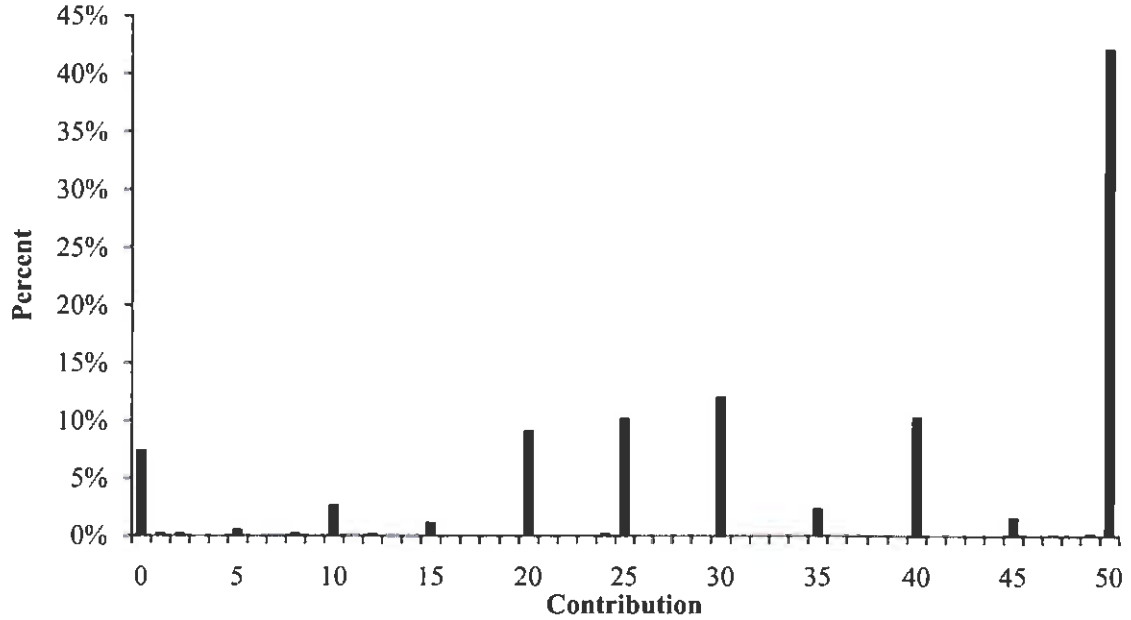
Note that the public good game is a strategic game where aspects besides from distributional preferences potentially affect behavior. For example, reciprocity might lead to participants being willing to contribute if they think that other group members contribute (this is often labeled conditional cooperation in the literature). For this reason, distributional preferences should not be expected to fully account for observed behavior.

Descriptive overview

The prediction of SET is that all group members keep their entire endowment and contribute nothing to the common pot. However, most public good experiments find that a substantial share of participants contribute a non-zero amount. Figure 5 shows that in our

sample, less than 8 percent contribute zero whereas the modal choice (about 42 percent of participants) is to contribute the full endowment of 50[14].

Figure 5: Contributions in the public goods game



Notes: The figure shows the distribution of contributions in the one-shot public good game. The average contribution is 35.2 and $n = 650$.

Taking distributional preferences into account yields predictions of behavior that, for some types, are different from the zero-contribution prediction of SET. We now loosely discuss these predictions before formalizing them below. First, participants with selfish preferences are in line with SET and maximize own income by contributing zero to the public good. The same holds for participants that are envious, spiteful or have kick-down preferences. On the other hand, for participants that are efficiency lovers it can be optimal to contribute a positive amount, irrespectively of what other group members do. Participants who have (strong) maximin preferences or are inequality averse should match the contributions of the other group members. For participants who have kiss-up preferences or are equality averse it can be optimal to both contribute more or less than the others, depending on the exact curvature of their indifference curves.

[14] These contributions are large compared to average observations in the standard lab. This is partly due to a subject pool effect of using a Danish sample. Hermann, Thöni and Gächter (2008) find in a cross-country study that subjects in Denmark are among the most cooperative.

Table 7 shows that, in accordance with the predictions, the average contribution of participants with efficiency concerns is greater than the overall average (38.5 vs. 35.2). The contributions of those who have maximin preferences or are inequality averse are strikingly close to the overall average (average of 35.6 and 35.2, respectively).[15] This is consistent with the prediction given rational expectations. Also in line with the predictions do those who have selfish preferences, those who are envious and those who are spiteful contribute less than the overall mean (31.6, 33.0 and 25.9, respectively), with the spiteful contributing the least of the three types. The equality averse is the on average least contributing type (28.6) while those with kiss-up preference contribute the most of all (41.3). The three participants that have kick-down preferences contribute more than the average (36.7). While the deviation from the mean is generally correct for all groups except for the kick-down type, participants still deviate from the predictions. One reason why participants with selfish preference on average choose to contribute a non-zero amount could be that intention-based concerns (e.g. reciprocity or guilt) matter as the actions of all group members affect outcomes.

***Table 7*:** Contributions to the public good

Preference type	Mean contribution	*N*
Efficiency lover	38.5	209
Inequality averse	35.2	151
Selfish	31.6	130
Maximin	35.6	89
Envious	33.0	36
Spiteful	25.9	17
Kiss-up	41.3	8
Equality averse	28.6	7
Kick-down	36.7	3
Overall	35.2	650

Notes: The table shows the mean contribution to the public good (between 0 and 50) for the full sample as well as split by preference types.

[15] The first wave of iLEE had several treatments relating to the public goods experiment. One treatment concerned the framing (*give* to the common pool vs. *take* from the common pool) and another concerned the incentives (paid vs. hypothetical). For details on the former treatment, see Fosgaard et al. (2010) and for the latter, see Tyran and Wengström (2009). In table 7, all treatments are merged but in the regression analysis below, we use dummies to control for any treatment effects.

Regression analysis

We now formally test whether the behavior across groups is statistically different. First, we introduce the notion of *weak* preferences. We say that preferences are weak if they are close to the selfish preferences in terms of the willingness to pay (to change the payoff of the recipient). In particular, we define weak preferences as having a positive WTP that is smaller than 0.32 or a negative WTP that is greater than -0.08. Graphically, this corresponds to the points that are adjacent to the center area (selfish) in figure 4.[16] We run tobit regressions with the contribution in the one-shot public game (*Contribution*) as the dependent variable, censored below at 0 and above at 50:

$$Contributi\,on_i = \beta_0 + \gamma \cdot Preference\,Types + \beta_1 \cdot Belief + \delta \cdot BasicContr\,ols + \varphi \cdot PsychContr\,ols + \varepsilon_i$$

As explanatory variables, we use dummies for the distributional preferences types (selfish is the reference category), participants' beliefs about other member's contributions (*Belief*) and two different sets of controls. We control for individual *Belief* as several papers have found a strong relation between beliefs and contributions[17]. The *Basic controls* variables are basic socio-economic background variables and include dummies for gender, education, and the participants' age and the age squared. It also includes dummies for the public good treatments (give vs. take framing and incentivized vs. hypothetical). The *Psych controls* include the participants' scores in each of the Big 5 personality traits and their number of correct answers to the IQ and cognitive reflection (CR) tests. Table 8 shows the results.

Model (1) in table 8 is the simplest specification where the contribution to the public good is explained only by the distributional preferences dummies. Model (2) adds the beliefs about other group members' contributions. Model (3) includes further the basic controls often applied in these settings and model (4) includes also the psychology controls. Overall, the table shows that weak preferences do not yield outcomes that are significantly different from the outcome under selfish preferences (in graphical terms, there is no difference between outcomes when indifference curves are vertically sloped and when they are close to vertical). This finding is as expected. The reason is that in our public goods game, the price of

[16] Our results are not depending on this classification. Whether we treat all participants in a category as the same or whether we make an even finer grid, we obtain qualitatively similar results.

[17] See for instance Thöni et al. (2010) for the sample at hand or Fischbacher and Gächter (2010).

increasing the payoff of others by 1 is 0.33 (a contributor loses 0.5 for every unit contributed, but the others collectively gain 1.5). Weak preferences imply that the WTP $\leq$ 0.32 and thus these participants should not be expected to contribute more than purely selfish ones.

Selfish participants contribute less than the average person and thus earn higher profits, i.e. they are in an advantageous domain. Using the selfish as a reference point, participants with maximin preferences, those that are efficiency loving and those that are inequality averse should increase their contribution if $WTP^a > 0.5/1.5 = 0.33$[18]. The table shows that for the non-weak preferences these three groups do in fact contribute significantly more than those that are selfish (the reference category in the regression). The contributions of participants with other types of preferences are not significantly different from the contributions of the selfish participants. This is potentially due to the low number of observations for each of these other types. The estimates are quite robust to the model specification and qualitatively hold for all model specifications, (1) to (4).

One particularly interesting aspect in the public good setting is the role of beliefs. Model (1) explains contributions with just the preference types as explanatory variables and finds the above mentioned effects. However, one could be worried that beliefs correlate with preference types. For instance, efficiency lovers might expect other group members to also have efficiency concerns and thus expect them to contribute more. Similarly, selfish participants might expect others to be selfish and thus to contribute nothing. As previous research has clearly shown that beliefs are highly correlated with contributions (see footnote 16), the preference dummies might just capture the effect of heterogeneous beliefs. Remarkably, however, we find highly significant effects of preference types even after controlling for beliefs. Hence, we do seem to be identifying the pure effect of underlying preferences on cooperation in the public good game.

[18] For the remainder of the paper, WTP^a refers to WTP in the domain of advantageous inequality and WTP^d to WTP in the domain of disadvantageous inequality.

Table 8: Distributional preferences and cooperation

Dependent variable: Contribution	(1)	(2)	(3)	(4)
Weakly efficiency loving	0.952	-0.208	-0.289	-0.663
	(1.527)	(1.474)	(1.706)	(1.707)
Efficiency loving	3.372***	2.613***	2.804***	3.043***
	(0.872)	(0.868)	(0.916)	(0.912)
Weakly inequality averse	-1.574	0.503	1.203	1.541
	(2.936)	(3.040)	(3.733)	(3.459)
Inequality averse	1.442*	1.833**	2.218**	2.007**
	(0.804)	(0.850)	(0.922)	(0.942)
Weakly maximin	-0.871	0.362	0.365	0.478
	(1.145)	(1.320)	(1.399)	(1.386)
Maximin	3.179***	2.270*	2.560**	2.095*
	(1.106)	(1.171)	(1.297)	(1.273)
Weakly envious	1.884	2.227	1.194	0.919
	(2.522)	(3.078)	(3.493)	(3.481)
Envious	0.393	1.474	1.665	1.908
	(1.385)	(1.272)	(1.340)	(1.404)
Spiteful	-2.144	-0.558	-0.549	-0.279
	(1.527)	(1.681)	(1.792)	(1.856)
Weakly kiss-up	2.904	1.704	3.037	2.030
	(1.938)	(2.790)	(3.379)	(3.622)
Kiss-up	3.449	1.403	1.489	-0.138
	(4.288)	(1.207)	(1.004)	(1.115)
Equality averse	-1.704	2.329	3.593	3.010
	(2.251)	(1.996)	(2.402)	(2.304)
Weakly kick-down	3.578	-0.318	2.235	0.945
	(5.277)	(6.515)	(6.694)	(6.015)
Belief		0.644***	0.744***	0.751***
		(0.036)	(0.045)	(0.045)
Basic controls	No	No	Yes	Yes
Psych controls	No	No	No	Yes
Log-pseudolikelihood	-1824.6	-1548.4	-1525.2	-1515.1
F-statistic	2.9	40.1	27.4	22.0
Prob > F	0.000	0.000	0.000	0.000
Pseudo R^2	0.010	0.160	0.172	0.178
N	650	650	650	650

Notes: The table shows marginal effects (at means) from tobit estimation of corporation. The dependent variable is *Contribution* in the one-shot public good game, double censored (below at 0 and above at 50). Explanatory variables include dummies for distributional preference types (*selfish* is the reference category). Weak preferences imply that indifference curves are (locally) close to vertically sloped. *Belief* is the expected average contribution by the three other group members. *Basic controls* are socio-economic variables and contain dummies for gender and education, a dummy for the *Give* framing, a dummy for hypothetical choices, the participants' *age* and *age squared*. *Psych controls* are participants' score in the Big 5 personality traits and their number of correct answers to the IQ and cognitive reflection tests. Parentheses contain robust standard errors. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Our estimates show that a participant with efficiency concerns on average contributes between Dkr. 2.6 and 3.4 more than a selfish participant depending on model specification.[19] The inequality averse contribute Dkr. 1.4-2.2 more than the selfish participants while for the participants with maximin preferences, the contribution is between Dkr. 2.1 and 3.2 larger. Selfish participants contribute on average Dkr. 31.6 (table 7). Hence, having efficiency concerns increases contributions by between 7.6 and 10.8 percent.

We now investigate whether leaving out the preference dummies affects the estimates for the other variables. Table 9 provides evidence on the effect of controlling for distributional preferences when explaining behavior in the public goods game. Model (5) is a basic model that is often employed when correlating corporation behavior with socio-economic variables and which does not contain the distributional preferences dummies. These are included in model (6) which is identical to model (3) in table 8. Model (7) includes the basic controls as well as the psychology controls but not distributional preferences. These are included in model (8) which is identical to model (4) in table 8. The general message from table 9 is that including the distributional preferences dummies does not alter the coefficients or significance levels of the other explanatory variables substantially, comparing (5) to (6) and (7) to (8). The reason for this is that the variables that correlate most with preference types (e.g. gender) do not correlate with corporation. Hence, we find no evidence that not controlling for distributional preferences biases the coefficients of other explanatory variables. Thus, in this particular setting, the most important reason for controlling for distributional preferences is to control for subject pool effects when comparing results across different samples.

[19] If we distinguish between those with a WTP smaller or greater than 1.20 (where the preferences of the latter must contain an element of altruism), we get an average marginal effect of 2.7 for those with WTP $\leq$ 1.20 and one of 3.1 for those with WTP > 1.20 (in model (2)). Both are different from zero (p = 0.005 and 0.041, respectively) but not different from each other (p = 0.815).

Table 9: The effect of controlling for distributional preference types

Dependent variable: Contribution	(5)	(6)	(7)	(8)
Belief	0.749***	0.744***	0.754***	0.751***
	(0.043)	(0.045)	(0.043)	(0.045)
Female	-0.162	-0.213	-0.639	-0.711
	(0.620)	(0.618)	(0.635)	(0.623)
Age	0.240*	0.239*	0.193	0.188
	(0.125)	(0.123)	(0.125)	(0.123)
Age squared	-0.003**	-0.003**	-0.002*	-0.002*
	(0.001)	(0.001)	(0.001)	(0.001)
Secondary education	-1.124	-0.621	-0.870	-0.344
	(1.229)	(1.245)	(1.312)	(1.323)
Short tertiary education	-1.423	-1.069	-1.492	-1.126
	(1.118)	(1.166)	(1.193)	(1.233)
Long tertiary education	-2.115*	-1.773	-2.382*	-2.009
	(1.241)	(1.289)	(1.323)	(1.362)
Give treatment	-4.512***	-4.573***	-4.397***	-4.441***
	(0.762)	(0.746)	(0.779)	(0.763)
Hypothetical treatment	-3.596**	-3.267**	-3.205**	-2.977**
	(1.433)	(1.406)	(1.454)	(1.417)
Agreeableness			0.190***	0.188***
			(0.060)	(0.060)
Conscientiousness			-0.042	-0.027
			(0.062)	(0.060)
Extraversion			-0.037	-0.057
			(0.057)	(0.058)
Neuroticism			-0.030	-0.026
			(0.053)	(0.053)
Openness			0.159***	0.167***
			(0.050)	(0.050)
IQ score			0.029	0.046
			(0.109)	(0.110)
CR score			-0.058	-0.213
			(0.314)	(0.323)
Preference types	No	Yes	No	Yes
Log-pseudolikelihood	-1533.2	-1525.2	-1523.0	-1515.1
F-statistic	63.2	27.4	37.5	22.0
Prob > F	0.000	0.000	0.000	0.000
Pseudo R^2	0.168	0.172	0.173	0.178
N	650	650	650	650

Notes: The table shows marginal effects (at means) from tobit estimation of corporation. The dependent variable is *Contribution* in the one-shot public good game, double censored (below at 0 and above at 50). Explanatory variables include *Belief* which is the expected average contribution by the three other group members, basic socio-economic variables (dummies for gender and education, a dummy for the *Give* framing, a dummy for hypothetical choices, the participants' *age* and *age squared*), psychological variables (the participants' score in the Big 5 personality traits and their number of correct answers to the IQ and cognitive reflection tests) and dummies for distributional preference types. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 9 also provides an opportunity to assess the relative effect of distributional preferences compared to other factors affecting behavior in the public good game. Several papers have found that the framing of the public good game is important for outcomes (e.g. Andreoni 1995, see Fosgaard et al. 2010 for an overview). In our setting, the average marginal effect of framing (*Give*) is -4.4 in model (8) while the effect of being efficiency maximizing is 3.0, i.e. almost 70% as big as framing. In addition, the effect of being efficiency maximizing is greater than the effect of making choices hypothetical[20]. Whether or not that is a lot is, of course, an open question. However, distributional preferences do seem to have an effect of a considerable size. This means that having more efficiency maximizers in the subject pool will significantly affect outcomes, highlighting the need to control for subject pool effects.

Predictive power of the behavioral model

Standard economic theory (SET) is a fairly bad predictor of behavior in the public good game as only 7.4 percent of subjects contribute zero to the public good. In fact, the modal choice is full contribution (42.2 percent of subjects). We now investigate how much of the gap between the SET prediction and the observed behavior can be explained by the behavioral model (BM) of distributional preferences. We find that accounting for distributional preferences explains almost half of this gap.

The starting point of this exercise is to formalize how distributional preferences translate into behavior. First, consider those who are concerned with efficiency. Participants with weak efficiency concerns have a $WTP^a \leq 0.32$. Thus, their willingness to pay is lower than the 0.33 that it costs to increase the income of the other group members by 1. Hence, they contribute zero. Analogously, efficiency concerned participants with $WTP^a > 0.32$ contribute the full endowment of 50.

Inequality averse participants and those with maximin preferences who are in the disadvantageous domain (i.e. they contribute more than they expect others to and thus earn less) reduce their contributions until they match the expected contributions of others. In the advantageous domain (where they contribute less and thus earn more) they increase their contributions if their $WTP^a > 0.32$. Thus, $c_i = Belief_i$ if $WTP^a > 0.32$ and $c_i = 0$ otherwise.

[20] See Tyran and Wengström (2009) for details and a discussion about the negative effect of hypothetical choices.

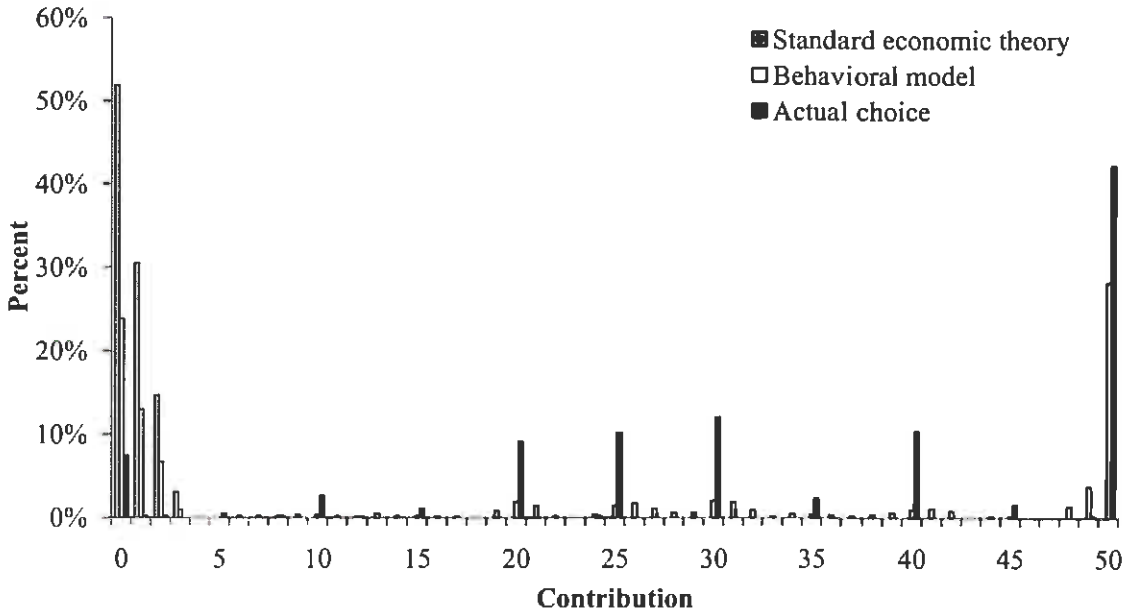
Selfish, envious, spiteful and participants with kick-down preferences all set $c_i = 0$. Participants with kiss-up preferences set $c_i = 0$ if $WTP^d < 0.32$ and $c_i = 50$ otherwise. Finally, equality averse participants maximize the difference between own earnings and the earnings of the other group members, i.e. they set $c_i = 0$ if $Belief_i > 25$ and $c_i = 50$ otherwise. The second column of table 10 summarizes these predictions.

We allow for participants to make mistakes by introducing a standard normal error term. Thus, the predictions of the behavioral model and standard economic theory are:

$$BM_i = \max(0, \min(c_i + \varepsilon_i, 50))$$
$$SET_i = \max(0, \min(0 + \varepsilon_i, 50))$$
$$\varepsilon_i \sim N(0,1)$$

We compare the observed choices with predicted behavior using both the behavioral model and the standard model for all decision makers. The black bars in figure 6 show the actual choices in the experiment (the same as in figure 5), the white bars show the prediction of BM and the grey bars show the prediction of SET. The figure hints that BM performs better than SET in explaining behavior as – for all levels of c_i where choices are observed – the height of the white bar is between the height of the grey and black bars.

Figure 6: Distribution of choices and predictions in the public goods game



Notes: The figure shows the distributions of actual choices (black bars) in the public good game as well as the prediction yielded by standard economic theory (grey bars) and the behavioral model implied by the distributional preferences (white bars). $N = 650$.

We formalize the analysis by calculating for each observation the absolute prediction error, i.e. the difference between the predicted and observed contributions. We do this for both predictions, BM and SET. Table 10 shows the average error for the two types as well as the share of the difference between behavior and SET that can be explained by BM for each preference type.

Table 10: Predictive power of the behavioral model

Preference type	Behavioral prediction (contribution = c_i)	Mean abs. error (SET)	Mean abs. error (BM)	Share explained by BM (%)
Efficiency loving	$c_i = 0$ if $WTP^a < 0.32$ $c_i = 50$ else	38.13	14.72	61.4
Inequality averse	$c_i = 0$ if $WTP^a < 0.32$ $c_i = Belief_i$ else	34.87	10.91	68.7
Selfish	$c_i = 0$	31.35	31.35	0.0
Maximin	$c_i = 0$ if $WTP^a < 0.32$ $c_i = Belief_i$ else	35.12	16.45	53.2
Envious	$c_i = 0$	32.69	32.69	0.0
Spiteful	$c_i = 0$	25.52	25.52	0.0
Kiss-up	$c_i = 0$ if $WTP^d < 0.32$ $c_i = 50$ else	40.45	33.29	17.7
Equality averse	$c_i = 0$ if $Belief_i > 25$ $c_i = 50$ else	28.25	30.63	-8.4
Kick-down	$c_i = 0$	35.20	35.20	0.0
Overall		34.88	19.17	43.11

Notes: The table shows how well observed choices are explained by both standard economic theory (SET) and the behavioral model (BM) based on the distributional preference types. The second column shows the how distributional preference types translate into predicted behavior. The third column shows the mean absolute error (observed behavior – (predicted behavior + error term)) for the SET model and the fourth column shows the corresponding number for the BM. The fifth column shows the how much of the gap between actual behavior and SET can be explained by taking distributional preferences into account, in the form of the BM.

Table 10 shows that the behavioral model performs better than the standard economic model for four of the nine preference types, whereas the opposite is true for one type (consisting of 6 participants) while the two models yield the same predictions for four preference types ($c_i = 0$). We measure the improvement of the BM as the share between the standard prediction and actual behavior that is explained by the BM. This improvement is between 18% and 69% for the four types where the behavioral model outperforms the standard model while it is minus 9% for the participants that are equality averse. The overall

weighted (by number of observations) average of the improvement is 43.1%. That is, the distributional preference types explain almost half of the gap between observed behavior and the predictions of standard economic theory. The behavioral model performs particularly well for those preference types whose optimal strategy it is to contribute the same (or more) as they expect others to contribute. One possible explanation for this finding could be that there is less scope for the influence of intention-based motives, such as reciprocity or guilt, when participants contribute the same as others. Hence, the assumption that only outcomes matter for decisions is more likely to be fulfilled and the test should perform better.

In summary, we find that pure distributional preferences do affect cooperation in strategic settings and agents who are efficiency maximizing, inequality averse or have maximin preferences contribute more than selfish people, even after controlling for beliefs. In addition, we find that almost half of the gap between observed behavior and behavior predicted by standard economic theory can be explained by taking distributional preferences into account.

5 Concluding remarks

Distributional (or social) preferences have been an important topic for economists in the past decade and there is increasing evidence supporting the hypothesis that agents care not only for their own material payoff but also for (some element of) the distribution of payoffs. The present paper adds to this mounting evidence.

In particular, we employ Kerschbamer's (2010) XY test to a large sample of the Danish population to make three contributions to the literature of distributional preferences. First, we show that role uncertainty does not affect behavior when compared with a treatment where roles are fixed ex ante. Second, we find that distributional preferences are heterogeneous in Denmark. In line with previous literature, we find that efficiency concerns are a more important driver for behavior than inequality aversion. The XY test employs a very comprehensive approach which allows for the full set of nine different preference types. We find four of these types to be of great importance empirically and together they account for 89 percent of subjects. In addition, we find that personal characteristics (socio-economic, attitudes and psychology measures) correlate with distributional preferences. Third, we find that distributional preferences influence corporative behavior in the public good game. In particular, we find that agents with non-selfish preferences contribute more to the public good

than those with selfish preferences and that the effects are substantial (contributions increase with between 6 and 11 percent), even after controlling for beliefs. Finally, we find that taking distributional preferences into account explains almost half of the difference between behavior and the prediction of standard economic theory. These findings can be interpreted as a validation of the XY test as it successfully predicts behavior in a different setting.

Taking distributional preferences into account to predict behavior in other games has not been widely done in the past. One exception is Fehr and Schmidt (1999) who investigate the predictive power of their model of inequality aversion in a series of games. Among these, they consider the last period of repeated public good games. They show that sustainable cooperation in the public good game is possible if sufficiently many group members are sufficiently inequality averse. However, given an assumption about the distribution of inequality aversion (based on evidence from ultimatum games), they show that these conditions will rarely be fulfilled. This is in line with their empirical evidence as contributions in repeated public good games is known to decline over the periods and little contribution is observed in the very last period (they consider several studies and find that on average 73 percent contribute zero in the last period). Note, however, that the final period of a repeated public good game is not directly comparable to a one-shot public good game as the one we conducted. While corporation is uncommon in the last period of a repeated game, it is not uncommon in one-shot games.

Another recent exception is Blanco, Engelmann and Normann (forthcoming) who estimate a model of inequality aversion and uses it to predict behavior in other games. They find that the prediction is fairly good on the average level but not on the individual level. While they look at behavior in the ultimatum game, the public good game and the sequential-move prisoners' dilemma game, we only consider behavior in the public good game. The predictive power of the XY test in other environments is a topic for further research.

We use the XY test because it is particularly simple and maps subjects into one of nine distinct, archetypical preference types. While other tests focus on the preference types that are expected to be most prevalent ex ante, the XY test can identify all forms of distributional preferences and we are thus able to verify which types are indeed most prevalent ex post. It turns out that of nine possible preference types, only four have strong relevance empirically. In addition, the XY test consists of a randomly ordered series of binary choices between two alternatives instead of an option to give or take from other subjects which is the normal procedure in the dictator game. As List (2007) shows, behavior in these game depends

crucially on the action-space of the subjects, i.e. whether they can only give or whether they can also take. By asking subjects to make simple left/right decisions, the XY test avoids this issue.

The XY test focuses purely on outcomes and not on intentions. That is, it does not take into account intention-based motives such as reciprocity or guilt aversion. These motives may influence behavior in settings where the decision maker does not solely determine the outcome, such as in the public good game. For example, an agent might have selfish preferences concerning the pure distribution of payoffs but at the same time be reciprocal. This could potentially explain why selfish agents do not contribute zero in the public goods game. Hence, when explaining behavior in such settings it might be beneficial to extent the XY model in order to incorporate both outcomes and intentions.

Using the internet as the platform for economic experiments is becoming increasingly popular as it has several advantages. Importantly for our investigation of the correlates of distributional preferences is that the internet allows us to get participants from all walks of life instead of just the standard student population often used in regular lab experiments. In addition, our corporation with Statistics Denmark leads to double blindness in our design. That is, participants remain anonymous to us throughout the experiment. This is important because a lack of anonymity leads to more pro-social behavior as highlighted by Levitt and List (2007). In addition, the anonymity of the internet ensures that participants consider the experiment truly one-shot as they will never face the subjects they are matched with. One drawback of using the internet is that it inevitably leads to a loss of control. For instance, subjects may not pay as much attention as in the lab which may explain the large share of inconsistent choices. However, given that we can identify and exclude these subjects, we feel that using the internet in this setting enables us to identify insights which would not have been able to do in the standard lab.

In conclusion, we find that distributional preferences are heterogeneous, they relate to personal characteristics and they influence behavior, thus making them important to control for.

References

Andreoni, J. (1995): Warm-Glow versus Cold-Prickle: The Effects of Positive and Negative Framing on Cooperation in Experiments. *Quarterly Journal of Economics* 110: 1-21.

Andreoni, J. and Miller. J. (2002): Giving according to GARP: An Experimental Test of the Consistency of Preferences for Altruism. *Econometrica* 70(2): 737-753.

Blanco, M., Engelmann, D. and Normann, H.-T. (forthcoming): A Within-Subject Analysis of Other-Regarding Preferences. *Games and Economic Behavior* forthcoming.

Bolton, G. E. and Ockenfels, A. (2000): ERC: A Theory of Equity, Reciprocity and Competition. *The American Economic Review* 90(1): 166-193.

Cappelen, A. W., Hole, A. D., Sørensen, E. Ø. and Tungodden, B. (2007): The Pluralism of Fairness Ideals: An Experimental Approach. *The American Economic Review* 97(3): 818-827.

Charness, G. and Rabin, M. (2002): Understanding Social Preferences with Simple Tests. *Quarterly Journal of Economics* 117(3): 817-869.

Costa, P. T. and McCrae, R. R. (2004): NEO PI-R. Manual – erhverv. *Dansk Psykologisk Forlag.*

Engelmann, D. and Strobel, M. (2004): Inequality Aversion, Efficiency, and Maximin Preferences in Simple Distribution Experiments. *The American Economic Review* 94(4): 857-69.

Engelmann, D. and Strobel, M. (2007): Preferences over Income Distributions: Experimental Evidence. *Public Finance Review* 35: 285-310.

Fehr, E., Fischbacher, U. (2002). Why social preferences matter: the impact of non-selfish motives on competition, cooperation, and incentives. *The Economic Journal* 112, C1–C33.

Fehr, E. and Schmidt, K. (1999): A Theory of Fairness, Competition, and Cooperation. *Quarterly Journal of Economics* 114: 817-868.

Fischbacher, U. and Gächter, S. (2010): Social Preferences, Beliefs and the Dynamics of Free Riding in Public Goods Experiments. *American Economic Review* 100(1): 541-556.

Fisman, R., Kariv, S. and Markovits, D. (2007): Individual Preferences for Giving. *The American Economic Review* 97(5): 1858-1876.

Frederick, S. (2005): Cognitive Reflection and Decision Making. *Journal of Economic Perspectives* 19: 25-42.

Fosgaard, T., Hansen, L. G. and Wengström, E. (2010): Framing and misperceptions in a Public Good Experiment. *Unpublished Working Paper.*

Herrmann, B., Thöni, C. and Gächter, S. (2008): Antisocial Punishment Across Societies. *Science* 319(5868): 1362-1367.

Holt, C. A. and Laury, S. K. (2002): Risk Aversion and Incentive Effects. *The American Economic Review* 92(5): 1644-1655.

Kerschbamer, R. (2010): The Geometry of Distributional Preferences and a Non-Parametric Identification Approach. *Unpublished Working Paper.*

Levitt, S. D. and List, J. A. (2007): What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World? *Journal of Economic Perspectives* 21(2): 153-174.

List, J. (2007): On the Interpretation of Giving in Dictator Games. *Journal of Political Economy* 115: 482-493.

Thöni, C., Tyran, J-R. and Wengström, E. (2010): Microfoundations of Social Capital. *Unpublished Working Paper.*

Tyran, J-R. and Wengström, E. (2009): Giving More When It's for Real: Evidence of a Negative Hypothetical Bias in Public Good Experiments. *Unpublished Working Paper.*

Varian, H. R. (1982): The Nonparametric Approach to Demand Analysis. *Econometrica* 50(4): 945-973.

Appendices

Appendix A: Instructions

Appendix B: Screen shots

Figure B1: Welcome

Figure B2: General instructions

Figure B3: Instructions, RandomRoles treatment

Figure B4: Choice

Figure B5: Confirmation

Figure B6: End of distributional preference test

Figure B7: Example of IQ puzzle

Appendix A: Instructions

The first part of the instructions is identical for both the FixedRoles and RandomRoles treatments (see figure B3 for a screen shot of the RandomRoles instructions, in Danish):

In this part of the experiment, there are two roles: decision makers and recipients. A **decision maker** makes 14 choices on behalf of the person him-/herself and a randomly selected second participant (**the recipient**).

Every choice is between two alternatives: LEFT and RIGHT. **The alternative chosen by the decision maker will determine the payment for both the decision maker and the recipient.**

Here is an example:

Vælg VENSTRE	VENSTRE		HØJRE		Vælg HØJRE
	Du får	Modtageren får	Du får	Modtageren får	
☐	70 kr	25 kr	50 kr	50 kr	☐

If the decision maker chooses LEFT, he/she gets 70 kr. and the recipient gets 25 kr. If the decision maker chooses RIGHT, he/she gets 50 kr. and the recipient gets 50 kr.

The continued instructions differ depending on the treatment:

Fixed role treatments:

Only decision makers are asked to make the 14 choices. Recipients make no decisions. Half the participants will be decision makers and the other half will be recipients.

What role you get is determined randomly before the decisions are made. It is as likely that you will be decision maker as it is that you will be recipient. Once the roles are determined, each decision maker is randomly matched with a recipient.

Only one of the decision maker's 14 choices will be selected for payment. All choices have the same probability of being selected.

On the next screen you will be told whether you have been chosen to be a decision maker or a recipient.

Remember that **if you are selected to be a decision maker**, your choices will determine **both** your own and a recipient's earnings from this part of the experiment. The recipient will only get a payment from your decisions and no further payment. **If you are selected to be a recipient**, your earnings will be solely determined by another participant's choices. In this case, you will not yourself make any choices in this part of the experiment.

Random roles treatment:

All participants are asked to make the 14 choices as if they are decision makers. Half the participants will actually be decision makers whose choices will count whereas the other half will be recipients whose choices will not count.

What role you get is determined randomly after the experiment has ended. It is as likely that you will be decision maker as it is that you will be recipient. Once the roles are determined, each decision maker is randomly matched with a recipient.

Only one of the decision maker's 14 choices will be selected for payment. All choices have the same probability of being selected.

On the next screens you will make the 14 choices between LEFT and RIGHT.

Remember that **if you are selected to be a decision maker**, your choices will determine **both** your own and a recipient's earnings from this part of the experiment. The recipient will only get a payment from your decisions and no further payment. **If you are selected to be a recipient**, your earnings will be solely determined by another participant's choices. In this case, your choices in this part of the experiment will have no effect on anyone's payment (neither on your own payment nor on anybody else's payment).

Subjects in the FixedRoles treatment see an additional screen informing them of the outcome of the random draw that determines their role:

Fixed Roles – Subjects chosen to be decision makers:

You have been randomly selected to be a **decision maker**.

On the next screens you will make the 14 choices between LEFT and RIGHT. Remember that your choices will determine **both** your own and a recipient's earnings from this part of the experiment. The recipient will only get a payment from your decisions and no further payment.

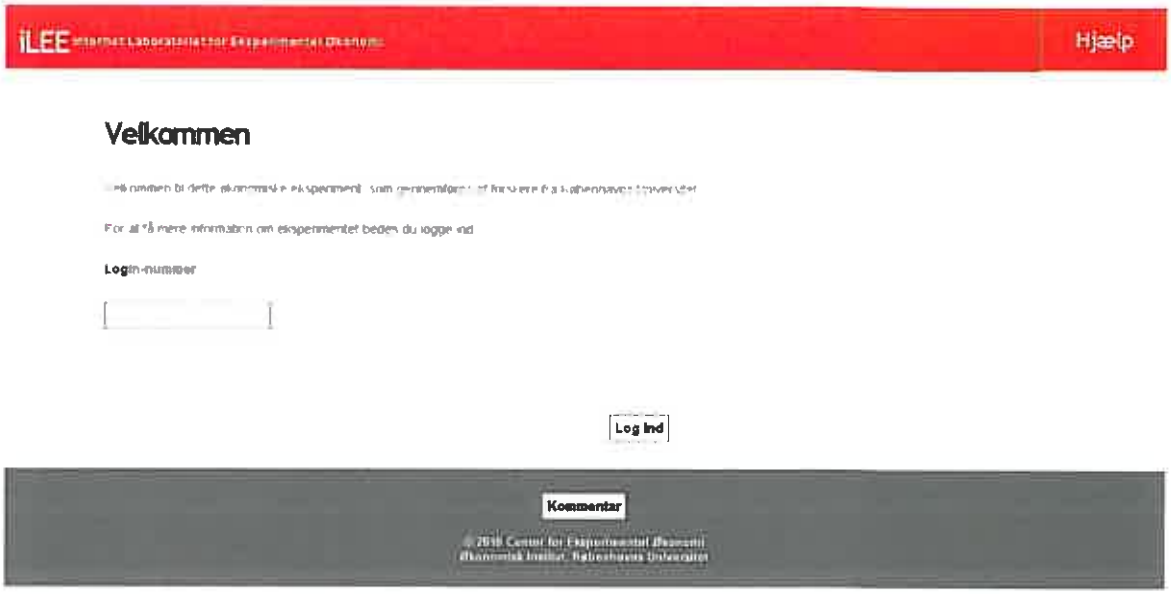
Fixed Roles – Subjects chosen to be recipients:

You have been randomly selected to be a **recipient**.

Your earnings will be solely determined by another participant's choices. You will not make choices in this part of the experiment yourself.

Appendix B: Screen shots

Figure B1: Welcome



Translation: **Welcome**.

Welcome to this economic experiment which is carried out by researchers from the University of Copenhagen.

To get more information about the experiment, please log in.

Login-number

Log in

Figure B2: General instructions

Translation: Information about the experiment.

You have now logged in. Thank you for your interest in the experiment.

Your participation is going to be valuable as you by completing the experiment contribute to Danish social sciences research.

In the experiment, you make money. The amount that you can earn depend on both your own and other participants' decisions. The amount is paid by bank transfer after you complete the entire experiment.

It is crucial for the scientific character of the experiment that you complete the entire experiment. **The experiment will last approximately 1 hour.** During the experiment, you can log out by closing your browser and you can return later before July 31st. When you log in again, you will continue from where you left.

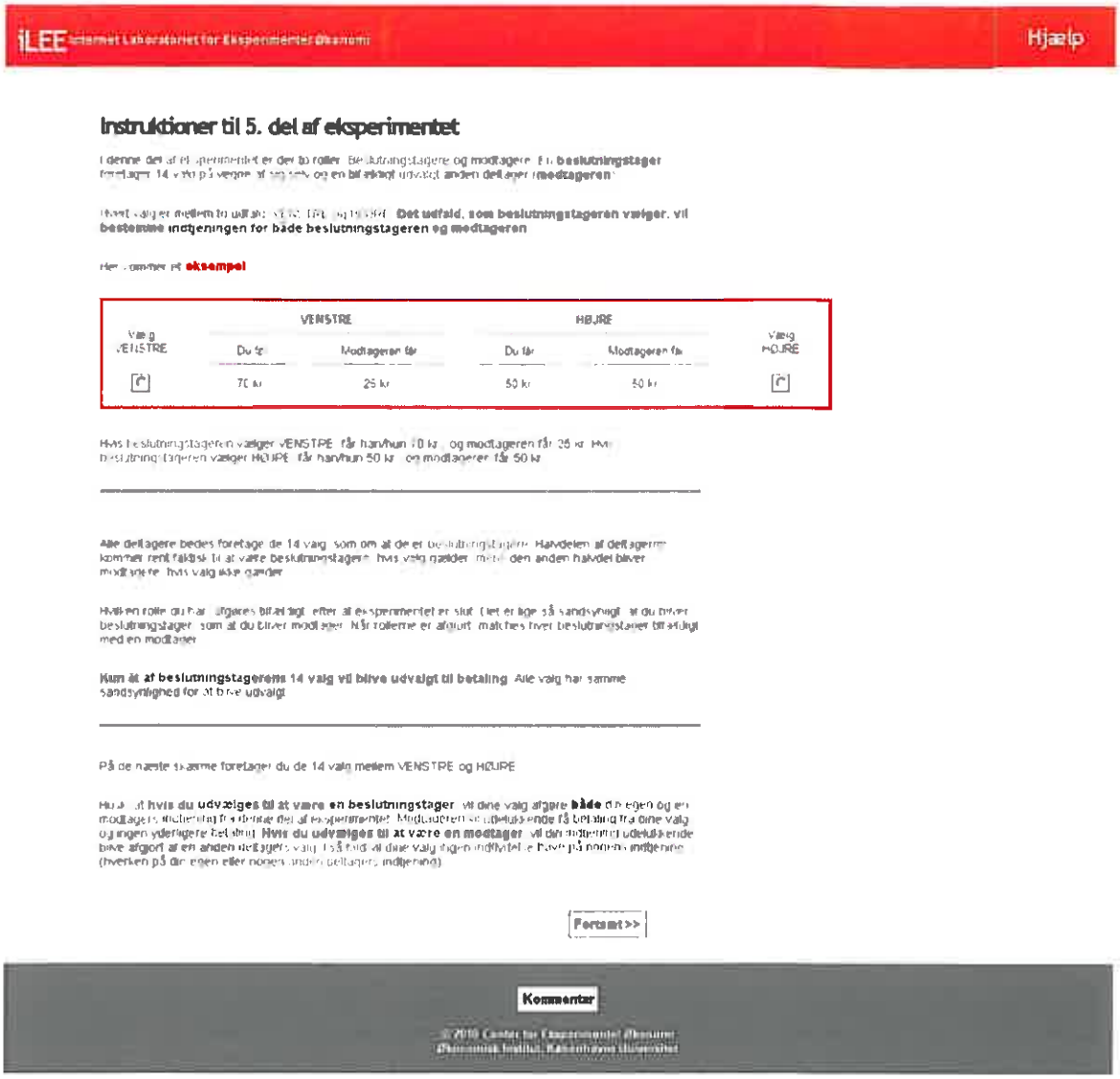
Like you, **the other participants in the experiment** were randomly selected by Statistics Denmark to receive a mailed invitation to participate in our first internet experiment which ran in May 2008 and like you, they all completed that experiment. Furthermore, they all, like you, were invited to participate in our second internet experiment, which ran in May-July 2009, and some (but not all) also completed this experiment.

Remember to save your login-number! You need it to log in again to see the outcomes of the experiment. Because you are anonymous, we cannot inform you of your log-in number in case you lose it.

If you have questions or need guidance, please send an email to ilee@econ.ku.dk or call telephone 35 32 44 04. You can find this contact information at any time by clicking the button Help (*Hjælp*) in the bar in the top of the screen.

Continue>>

Figure B3: Instructions, RandomRoles



Translation:

This screen shows the instructions for the RandomRoles treatment. See appendix A for full translation of the instructions for both treatments (RandomRoles and FixedRoles).

Figure B4: Choice

Translation: **Choice (1/14)**

Select your preferred outcome

Vælg VENSTRE = Select LEFT.

Du får = You get.

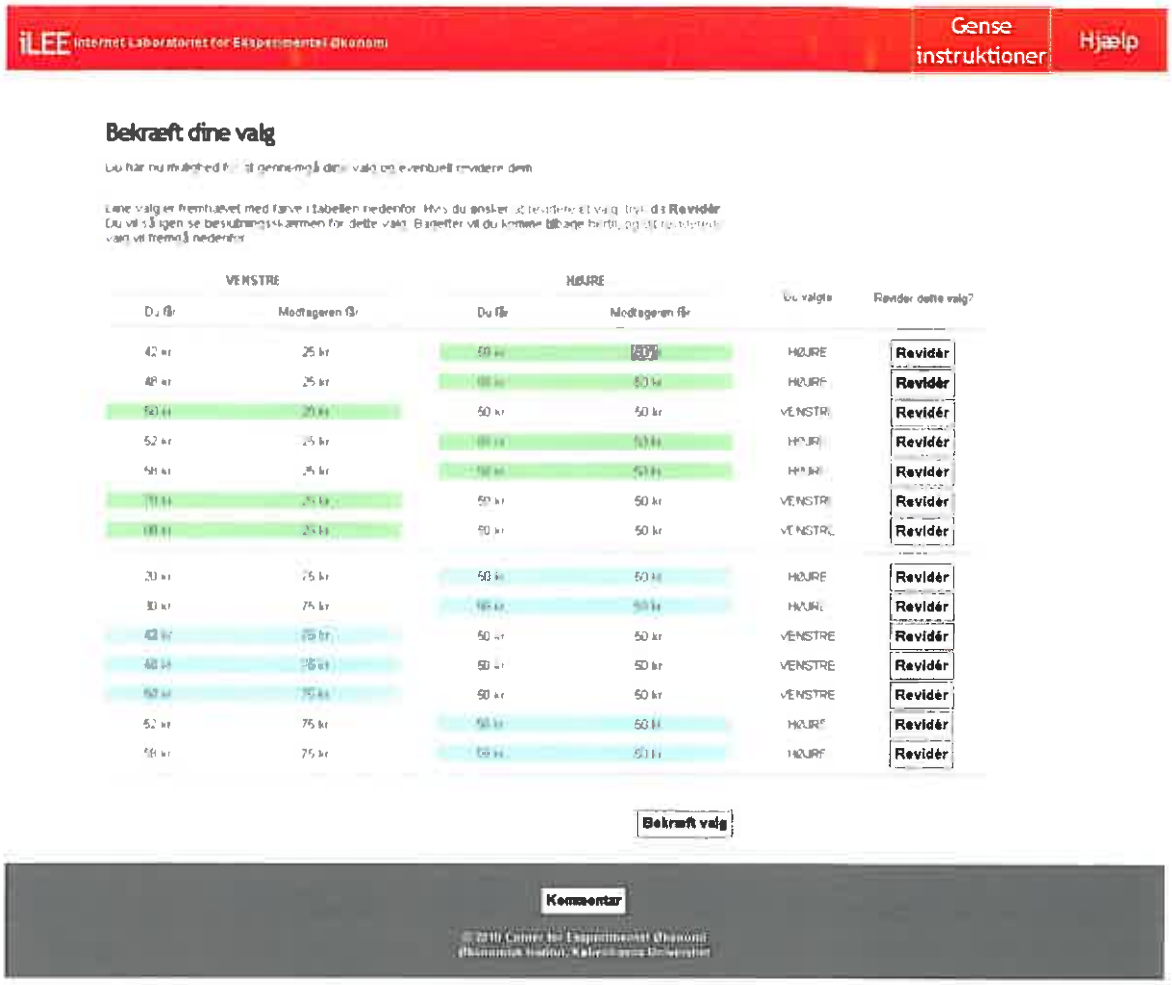
Modtageren får = The recipient gets.

Vælg HØJRE = Select RIGHT.

Indsend svar = Send answer.

Top bar: Gense instruktioner = Repeat instructions, Hjælp = Help

Figure B5: Confirmation



Translation: **Confirm your choices**

You now have the option to examine your choices and possibly to revise them.

Your selections are pointed out by colors in the table below. If you wish to revise a decision, click **Revise** (Revidér). You will then again see the decision screen for this decision. Afterwards you will return here and your revised choice will be apparent below.

VENSTRE = LEFT, HØJRE = RIGHT

Du får = you get, modtageren får = the recipient gets.

Du valgte = You chose, Revidér dette valg? = Revise this decision.

Bekræft valg = Confirm decisions

Top bar: Gense instruktioner = Repeat instructions, Hjælp = Help

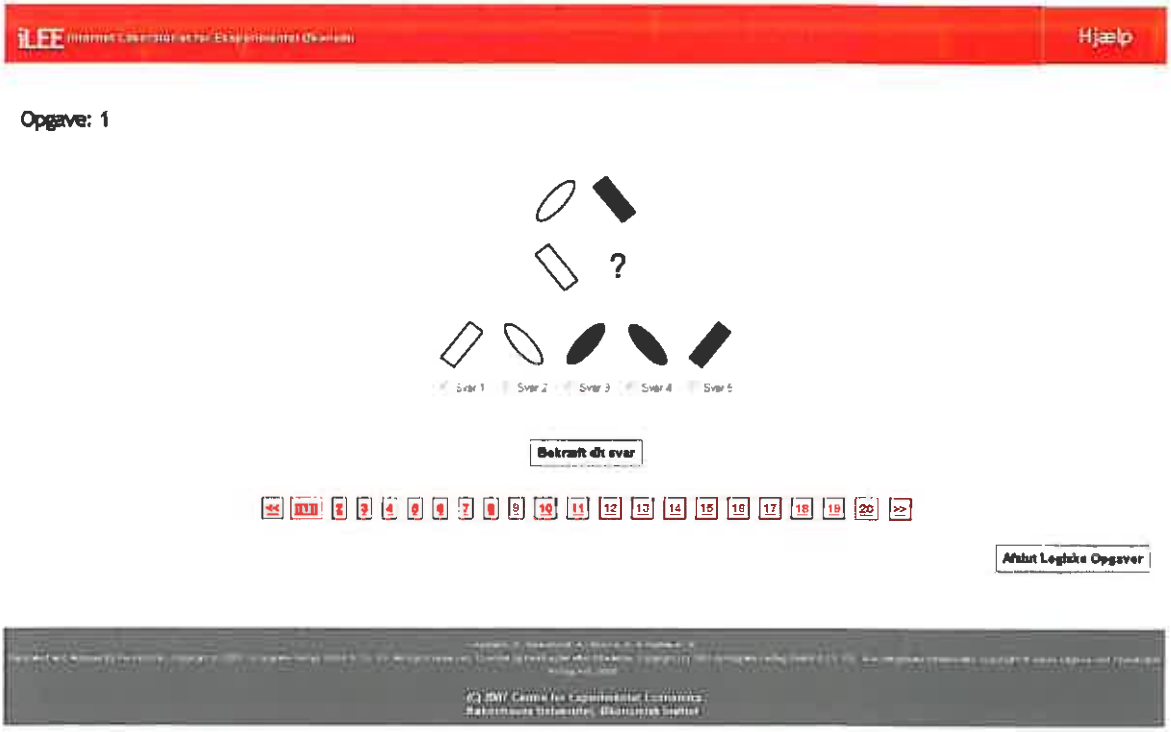
Figure B6: End of distributional preference test

Translation: **The fifth part of the experiment is now over.**

You will be told the outcome once the entire experiment is over.

Continue

Figure B7: Example of IQ puzzle



Translation: **Puzzle 1**

[Picture]

O Answer 1 O Answer 2 O Answer 3 O Answer 4 O Answer 5

Confirm your answer

<< 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 >>

End logic puzzles

Chapter 3

To See is to Believe: Common Expectations in Experimental Asset Markets

Stephen L. Cheung, Morten Hedegaard and Stefan Palan

To See is to Believe: Common Expectations in Experimental Asset Markets

Stephen L. Cheung, Morten Hedegaard and Stefan Palan*

March 2011

Experimental asset markets of the type introduced by Smith, Suchanek, and Williams (1988) are known to produce price bubbles and crashes with inexperienced subjects. This study investigates the conjecture that this phenomenon may be explained in part by the fact that traders are uncertain about the behavior of others. In particular, we analyze the effect of manipulating common expectations in a group of individually rational traders. Individual rationality is induced by requiring participants to correctly answer an extensive set of control questions. With this in place, common expectations is then manipulated by varying the knowledge that traders have regarding the fact that all others in the market are also required to answer these questions. We find that markets in which this common knowledge is absent do not differ significantly from baseline markets in which traders do not answer control questions. However, when it is common knowledge that all must answer the questions correctly, we find that mispricing is essentially eliminated.

Keywords: asset market experiment, price bubbles, common knowledge of rationality.

JEL codes: C92, D84, G12.

---

* Cheung: School of Economics, The University of Sydney, Merewether Building H04, Sydney NSW 2006. Stephen.Cheung@sydney.edu.au
Hedegaard: University of Copenhagen, Department of Economics, Øster Farimagsgade 5, building 26, DK-1353 Copenhagen K. Morten.Hedegaard@econ.ku.dk.
Palan: Institute of Banking and Finance, Karl-Franzens University Graz, Universitätsstraße 15/F2, 8010 Graz. Stefan.Palan@uni-graz.at
We thank the Centre for Experimental Economics at the University of Copenhagen for access to their laboratory and subject pool, and our respective universities for financial support. We thank Charles Noussair, Robert Slonim and Jean-Robert Tyran for valuable discussions. An earlier draft of this paper circulated under the title "Complexity, confusion and bubbles in experimental asset markets".

Introduction

In asset market experiments of the type pioneered by Smith, Suchanek, and Williams (1988), hereinafter SSW, it is well-known that price bubbles – "trade in high volume at prices that are considerably at variance from fundamental value" King et al. (1993) – followed by crashes are commonplace when subjects are inexperienced. In the two decades since SSW first documented this pattern, an extensive body of literature, much of which is summarized in Porter and Smith (2008), has sought to identify the cause of these bubbles by applying a very wide range of manipulations to the market environment. These efforts have largely proven fruitless, such that the conventional wisdom remains that the only condition known to eliminate these bubbles is repeated experience in a stationary market environment, as part of the same group of subjects.

The behavioral pattern of bubbles and crashes is a puzzling deviation from the prediction of economic theory: in an environment in which the fundamental value of the asset being traded is common knowledge, any trade that occurs should take place at prices exactly equal to fundamental value, under the assumptions of risk neutrality and common knowledge of rationality. Porter and Smith (1995) examine the role of risk preferences by eliminating uncertainty over the realizations of dividends. They find that price bubbles still occur when dividends are certain and conclude from this that risk preferences are not the cause of mispricing. Common knowledge of rationality has, until now, remained untested. This is the topic of the present paper.

In their seminal paper, SSW hypothesize that a lack of common expectations among subjects might be the cause of mispricing[1]. Even if all traders are rational, some might believe that others are not. In this case, speculative behavior can be optimal. Smith (1994) note that simply reading out instructions to subjects is no guarantee that the subjects will have common expectations, since there may still be uncertainty as to how each subject will use the information.

In addition to reading the instructions out loud, we examine the effect of requiring all subjects in a market to correctly answer an extensive series of control questions before the

[1] "What we learn from the particular experiments reported here is that a common dividend, and common knowledge thereof is insufficient to induce initial common expectations. As we interpret it this is due to agent uncertainty about the behavior of others." (Smith, Suchanek, and Williams (1988)).

experiment commences. We believe that this series of questions ensures that our subjects understand the environment and, thus, improves rationality at the individual level. In our "common knowledge" treatment (CK), all ten traders in a market must answer all questions correctly and this is common knowledge. That is, we tell subjects that the experiment will only begin after everyone has answered all of the questions correctly. In the "no common knowledge" treatment (NCK), all traders in the market must still answer all questions correctly but this is not common knowledge. In particular, they only know that ten out of twenty subjects in the session are required to answer the questions – what they do not know is that all ten are in the same market. The other ten subjects in the session are not required to answer any control questions, and are simply told to wait until the experiment begins.

Thus, the *individual* understanding induced by requiring all subjects in a market to answer control questions is the same in our NCK markets as it is under CK. All that differs between these treatments is whether or not the subjects know for sure that the other traders in their market are also required to answer the questions. Hence, our treatments allow us to identify the effect of common expectations and, thereby, to test the original hypothesis of SSW.

We find that CK essentially eliminates bubbles in four out of six markets and leads to less mispricing overall. In our NCK markets in which all traders must also answer the control questions correctly – but where this is not common knowledge – we find substantial mispricing similar to standard SSW markets. Thus, the fact that it is not common knowledge that all traders in a market are required to answer the control questions is responsible for the greater mispricing in our NCK treatment as compared to CK. This is in support of the conjecture of SSW that the lack of common expectations about other subjects' behavior is a major source of mispricing in experimental asset markets of this type.

The paper is organized as follows. Section 2 briefly reviews the most pertinent previous research. Section 3 describes our design and procedures, section 4 presents our results, and section 5 concludes.

2 Experimental asset markets

The canonical design of an experimental asset market is due to SSW.[2] In this design each market typically has between eight and twelve traders, each of whom is given an initial endowment of experimental money and shares which they can trade in a computerized double auction. The market operates over fifteen trading periods, with each period typically running for four minutes.

At the end of each period, each share pays a dividend to its current owner. This dividend is the same for all shares, and the probability distribution from which it is drawn is common knowledge. In particular, the dividend takes values of 0, 8, 28 or 60 currency units, each with equal probability, such that the expected dividend in each period is 24. After the fifteenth dividend has been paid, shares have no remaining value. The fundamental value of each share is thus 24 times the number of remaining dividend payments, so it is 360 in the first period and declines by 24 after each dividend has been paid. Not only are these facts common knowledge, they are also typically presented to subjects in the form of an "average holding value table". Table 1 shows an example of such a table, as first introduced by Porter and Smith (1995).

[2] The following description relates to design 4 in SSW, which is the standard and most extensively-studied set of parameters.

Table 1: Average Holding Value Table

End Period	Begin Period	Periods Held	×	Average per Period Dividend Value	=	Average per Unit Inventory Value
15	1	15	×	24	=	360
15	2	14	×	24	=	336
15	3	13	×	24	=	312
15	4	12	×	24	=	288
15	5	11	×	24	=	264
15	6	10	×	24	=	240
15	7	9	×	24	=	216
15	8	8	×	24	=	192
15	9	7	×	24	=	168
15	10	6	×	24	=	144
15	11	5	×	24	=	120
15	12	4	×	24	=	96
15	13	3	×	24	=	72
15	14	2	×	24	=	48
15	15	1	×	24	=	24

Notes: This table shows the "average holding value table" provided to subjects by Porter and Smith (1995) as part of the experimental instructions. The last column shows, for each period in the second column, the fundamental value of the asset.

By designing this relatively simple market environment, SSW originally set out to create:

> "a 'transparent' asset trading market where shares had a well-defined intrinsic value based on common trader information concerning share expected, or average, dividend value. Using these experimental results as a baseline, the research program originally was expected to inquire if price bubbles – trading away from intrinsic value – could be created by controlling information or other elements." (Porter and Smith (2008))

Contrary to expectation, this intentionally simple design has instead been found to consistently generate price bubbles and crashes when subjects are inexperienced. It is typical for these markets to exhibit substantial price deviations from fundamental value, starting out below fundamental value in the first few periods before rising above it – sometimes to a level in excess of the maximum possible dividend value of a share –before crashing back down toward fundamental value as the end of the market approaches.

One explanation for why initial trades typically occur at prices below fundamental value is that they could be motivated by risk aversion on the part of sellers. On the other hand, the only *rational* explanation for purchases at prices exceeding the maximum dividend value is that these trades are motivated by speculation. That is, the only reason why a rational trader would be willing to pay such a price would be if she believed there was a second trader to whom she could resell at an even higher price – and for that to be the case, this second trader would have to either himself be irrational or else believe that there was some third trader to whom he could resell at a yet higher price, and so on.[3]

Building upon such reasoning, it can be argued that even though the *dividend process is common knowledge*, this is insufficient to induce *common knowledge of rationality*. For this reason, a price bubble can occur even when all traders are indeed rational and correctly understand the dividend structure of the asset – but there are at least some traders who believe (that some others believe, ...) that some others may not. Alternatively, and more directly, it could simply be the case that overpriced transactions are the result of actual irrationality (or confusion, or decision errors) on the part of some traders.

The hypothesis that price bubbles are due to failure of common knowledge of rationality is consistent with the observation that prices track more closely to fundamental value as subjects gain repeated experience in the same market environment as part of the same group of traders (SSW; van Boening, Williams, and LaMaster (1993)). Nonetheless, it is unclear from this result to what extent it is actual rationality, as opposed to the common knowledge thereof, that is improved with experience.

Conclusive evidence of *actual irrationality* in inexperienced markets is provided by Lei, Noussair, and Plott (2001), who control for speculative motives by assigning each subject to a role either as a buyer (with no opportunity to resell) or seller (with no opportunity to repurchase). They find that 38 percent of all trades in such markets occur at prices in excess of the *maximum* dividend value where, given that it is not possible to resell, the buyer is sure to incur a net loss from such a trade.[4]

---

[3] Of course, trades at less extreme prices might also be motivated by speculation. Moreover, speculation may still be profitable even when the resale price is lower than the purchase price, since the speculator can earn dividends during the time that she holds the share.

[4] In the (Lei, Noussair, and Plott) no-speculation sessions the dividend is either 20 or 40 with equal probability, such that the maximum dividend value is 4/3 times fundamental value. By contrast, in the SSW design the

Consistent with the results of Lei, Noussair, and Plott (2001), in markets in which subjects are inexperienced the bubble-and-crash pattern has proven highly robust to a very wide range of treatment manipulations involving various aspects of the market environment. Much of this work is surveyed in Porter and Smith (2008).

One conspicuous exception is provided by Noussair and Tucker (2006) who find that when a complete set of futures markets is provided, price bubbles in the spot market are eliminated. In their design, these futures markets have the distinctive feature that they are opened sequentially in reverse order – starting with the futures market for period fifteen – *before* the spot market is opened. One implication of this procedure is that fully half of the session time is taken up with this sequential opening of futures markets before the first period of spot trading commences. Noussair and Tucker (2006) explicitly acknowledge that this is done to "to facilitate the backward reasoning that is required for agents to realize that the expected future dividend stream corresponds to a limit price for a rational trader" (p. 172).

Noussair and Tucker (2006) report one of the very few treatments to eliminate the bubble-and-crash phenomenon for inexperienced subjects in the SSW environment.[5] However, it remains unresolved whether their result holds because they provide a complete set of markets (as might be suggested by theory) or, as they themselves suggest, because their treatment hammers home the logic of backward induction to subjects. Even if it is the latter, it is not possible in their design to disentangle whether the diminished mispricing they observe is the direct effect of training subjects in the logic of backward induction, or a byproduct of the fact that it is common knowledge that all subjects have undergone this training.

As we describe below, our control questions are also designed to train subjects to use the logic of backward induction to calculate the fundamental value of the experimental asset.

---

maximum dividend value is 60/24 = 2.5 times fundamental value. This difference contributes to the relatively high incidence of overpriced transactions in (Lei, Noussair, and Plott) data.

[5] Subsequent to commencing the work reported here, we also became aware of the results of (Lei and Vesely (2009)). In their experiment there is a pre-market phase during which participants passively experience the realisation and accrual of a stream of dividends at periodic intervals, with each interval corresponding temporally to the length of a trading period. We observe that this procedure is similar to that of (Noussair and Tucker (2006)) insofar as, prior to the commencement of trade in the asset market proper, participants first gain experience in a related activity for a length of time equal to the life of the experimental asset itself. Both results may thus also be related to the robust finding, noted above, that price bubbles tend to be diminished with repeated experience as part of the same group of participants.

Since our design does not involve futures markets, we can eliminate complete markets as an explanation for any effect that we observe. Moreover, our design enables us to directly manipulate whether or not it is common knowledge that all traders in a market have undergone the training.

As it turns out, this control proves to be critical. We find that our control questions alone do not have any significant effect upon the severity of mispricing. However, when it is common knowledge that all traders in a market must complete the questions successfully, we find that mispricing is significantly diminished.

3 Design and Procedures

In designing our experiment to address the issue of subjects' divergent expectations, we follow the recommendations of Haruvy and Noussair (2006). They argue for a design which (i) facilitates subjects' comprehension of the link between the expected future dividend stream and the fundamental value of the asset, (ii) ensures that subjects are not trading due to an experimental demand effect, and (iii) induces common knowledge that subjects are using the expected future dividend stream as their limit price.

We address issue (i) by requiring subjects to successfully complete an extended set of control questions. Whether or not it is common knowledge that all traders in a market must answer these questions depends upon the treatment, and thus our research question relates directly to issue (iii). As regards issue (ii), one might fear that by their very nature, our control questions (which we describe below) might directly induce our subjects to trade at fundamental value. If this was the case, any effect that we observe might then be attributable to an experimental demand effect. However, given that our control questions are identical in our CK and NCK treatments, any demand effect should also be identical across these treatments, and hence should not affect a between-groups comparison of these treatments. As it turns out, our results reveal minimal effects of the control questions alone in the absence of common knowledge. For this reason, we are confident that our results are not contaminated by any significant demand effects.

Depending upon the treatment, our control questions may achieve one or both of the following two effects:

Controlling for confusion on the part of subjects (treatments CK and NCK). Although it is standard procedure in SSW-type markets to carefully explain the dividend process and to provide subjects with an average holding value table similar to Table 1, little is known regarding the extent to which subjects either understand the information or make use of the table. In particular, in contrast to some other branches of experimental economics research, the literature on asset market experiments following SSW does not consistently make use of control questions to check on subjects' understanding of the decision environment. However, given the evidence of confusion or decision errors identified by Lei, Noussair, and Plott (2001), we seek to control for such misunderstandings by implementing a more robust structure of control questions than is typical in this literature.

Improving common knowledge of rationality (treatment CK). As discussed above, price deviations from fundamental value may arise from uncertainty concerning the expectations of other traders. In our CK treatment we make it common knowledge that all subjects in the market must successfully answer all of the control questions, thereby strengthening the common expectation that all subjects have understood the fundamental value process and that they will behave accordingly.

Basic environment and endowments

We adopt the classic SSW parameters of fifteen double auction trading periods each lasting four minutes. Each share pays a dividend of 0, 8, 28 or 60 currency units, each with equal probability, at the end of each trading period. Subjects are randomly assigned to receive one of three different initial endowments of experimental currency and shares, all having the same fundamental value. These endowments are summarized in Table 2.

Table 2: Endowments and Exchange Rates

Endowment type	I	II	III
Number of traders of this type	3	4	3
Initial stock	2	4	6
Initial cash	1,890	1,170	450
Endowment value (ECU)		2,610	
Exchange rate (DKK/ECU)		1/11	
Endowment value (DKK)		237.27	
Total Stock of Units (TSU)		40	

Notes: The table shows the technical parameters of the experimental design. One DKK is approximately equal to 0.20 USD or 0.13 EUR (as of November 2009).

Control questions

Prior to constructing our control questions, we conducted a thorough search of the literature on experimental asset markets for appropriate precedents, and identified very few. Given that control questions are typically only included in working papers and do not find their way into final publications, we cannot claim on these grounds that they are seldom used, but it is nonetheless clear to us that their use is not universal. Moreover, many of the examples we identify relate to aspects of the market that are novel to a specific paper (for example, the futures market treatment of Noussair and Tucker (2006)), as opposed to the standard SSW environment itself. In short, the existing literature provides little clear guidance as to what to include in an appropriate set of control questions.

Our main intervention involves two sets of control questions – one framed from the perspective of buying a share, and the second framed from the perspective of selling. Since one aspect of our interest in control questions is to train subjects in the logic of backward induction without providing a complete set of futures markets, we include fifteen questions in each frame, ordered from period fifteen to period one.[6] For example, the first buyer control question asks:

> Suppose that you buy one share in period 15 and that you keep it until the end of the market (i.e. until period 15). What is the average total dividend that you will receive from this share?

Similarly, the second seller control question asks:

> Suppose that you sell one share in period 14 and that you do not buy it back. What is the average total dividend that you give up on this share?

Our control questions thus effectively require each subject to enter the values from the final column of the average holding value table twice, from the bottom up. The majority of subjects learn to do this relatively quickly, and without requiring any assistance from the experimenters. However, in each session there are also up to 20 percent of subjects who take much longer – in some cases over twenty minutes – requiring further instruction from an experimenter in the process.

---

[6] In addition, all participants were also required to answer a set of four basic control questions. See Appendix A for details.

Details of sessions

Our experiments were conducted at the Laboratory for Experimental Economics at the University of Copenhagen between October 2009 and June 2010. We oversubscribed sessions to ensure that there would be exactly ten subjects in each market. We also operated two completely independent markets in each session, for a total of twenty subjects. No subject had ever taken part in any previous asset market experiment. Each session lasted up to 2.5 hours, and the average earnings were 230 Danish kroner (approximately 46 US dollars, 31 Euros as of November 2009). The experiments were conducted in English, and the computerized market was programmed using the z-Tree environment (Fischbacher (2007)).

At the start of each session, we first distributed and read aloud the first three pages of instructions which deal with the mechanics of using the computer interface to make price offers and to buy and sell shares.[7] This was followed by a ten-minute practice period, which did not count toward subjects' earnings. Note that subjects completed this practice task before they had been told about the dividend structure of the asset or how their earnings would be determined. We next circulated and read aloud the remainder of the instructions, dealing with the dividends, average holding value table and calculation of earnings. Following this, we required subjects to complete the control questions described above, as appropriate to the treatment (as discussed below). Upon conclusion of the experiment, we also asked subjects to complete a questionnaire.

Treatments

Our experiment consists of four treatments, referred to as Baseline, CK, NCK, and WAIT. The first three form the main focus of our analysis, while WAIT is a necessary byproduct of the procedures we use to obtain the NCK treatment.

The Baseline treatment is a standard SSW market with the parameters outlined above. Subjects in the Baseline treatment only answer a set of four basic control questions (see Appendix A) and not the extended set of thirty buyer and seller control questions. This treatment provides a benchmark for the severity of mispricing in SSW markets conducted under standard procedures, as applied to our subject pool.

In the Common Knowledge (CK) treatment all subjects must correctly answer both the four basic control questions and the extended set of thirty buyer and seller control questions.

---

[7] The full text of the instructions for the CK treatment is contained in Appendix C of our working paper.

Subjects are explicitly informed that the experiment does not begin until all twenty subjects in the session have correctly answered all of these questions. This design is intended to induce common knowledge of rationality, in the sense that all subjects have understood the instructions, that all subjects believe that all subjects have understood the instructions and that all subjects believe that all subjects believe that all subjects have understood the instructions, etc. In support of this, as will be seen in our results below, it appears that the CK design very quickly creates common expectations (typically within the first two periods) about other subjects' behavior in the market.

To obtain our last two treatments, we inform all twenty subjects in a session that an unspecified number of them will be asked to answer a set of control questions, and that these subjects must answer all of the questions correctly before the experiment can begin. The remaining subjects will not be asked to answer any questions, and must simply wait until the experiment begins.

Of the twenty subjects in the session, we require ten to answer the full set of control questions. Through a message on their computer screens, we inform these ten subjects that exactly ten of the twenty subjects in the session are being asked to answer control questions. What they are not told is that we then group these ten subjects together to trade in the same market. This market thus contains ten subjects who have all answered the control questions correctly (with the implied effect upon their understanding of the instructions) but who are not aware that all others in the market are also required to answer these questions. In other words, we induce individual rationality (as in CK) but not the common knowledge of having done so. We refer to this third treatment as No Common Knowledge (NCK).

As a byproduct of our NCK treatment we also have ten subjects in each session who do not answer any control questions and must simply wait for the others to finish. These ten subjects are also grouped together to make up the second market in the session. Through a message on their computer screens, we inform these subjects that when the experiment begins, none of the subjects in their market will have answered control questions. That is, the unspecified number of subjects who must answer the questions will be in a different market to them. We refer to the fourth treatment as WAIT. This treatment is similar in design to the Baseline, except that subjects must wait approximately twenty minutes before the experiment begins. Compared to Baseline, the WAIT institution may thus give subjects additional time to think.

We collect six independent observations (markets) in each of our treatments. In the analysis below, each market is treated as the (independent) unit of observation.

4 Results

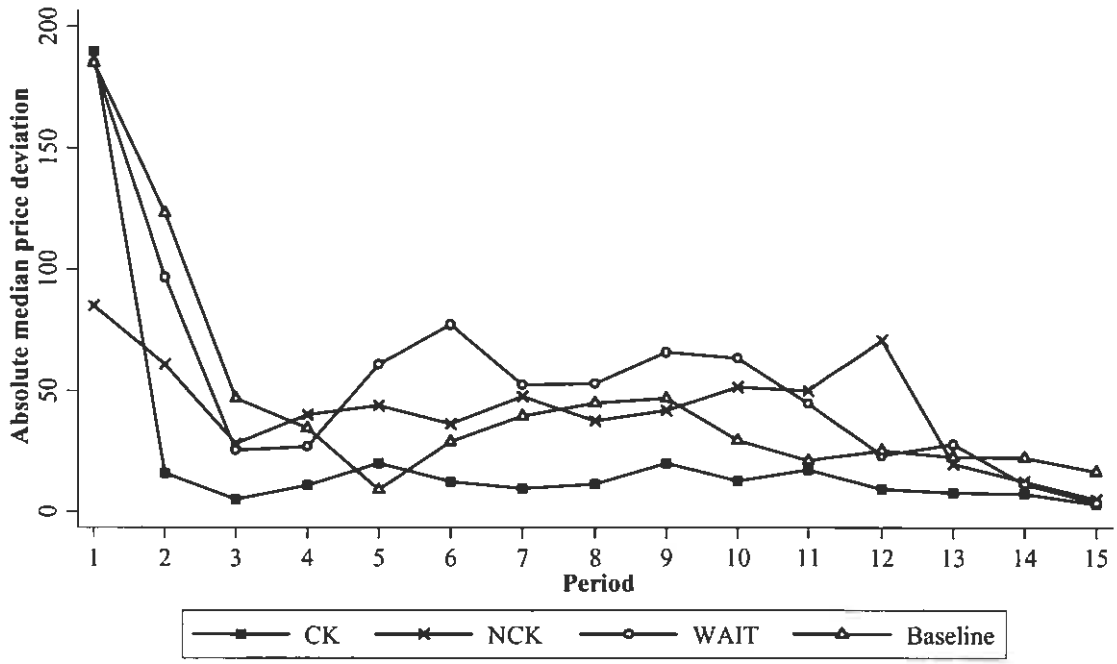
We begin this section by providing a descriptive overview of our results, before turning to a formal analysis of bubble measures and regressions. We find that control questions alone do not have a great effect upon mispricing, but that the combination of control questions together with common knowledge essentially eliminates mispricing in most markets.

4.1 Descriptive overview

Figure summarizes our main findings by showing aggregated measures of mispricing for each of our four treatments. For each market and in each period, we calculate the absolute deviation of the median transaction price from fundamental value. For each treatment and in each period, we then calculate the median over the six markets and plot the result in figure . Through simple visual inspection, we see that mispricing is lowest in the CK treatment in thirteen of the fifteen periods. There is no obvious ranking between the NCK, WAIT and Baseline markets. If anything, WAIT appears to show the greatest mispricing in the middle periods, while the picture is mixed in the earlier and later periods.[8]

[8] We ran a fifth treatment that was identical to CK but assigned fixed buyer and seller roles as in (Lei, Noussair, and Plott (2001)) and used a certain dividend as in (Porter and Smith (1995)). The results were comparable to those of CK, in that prices closely tracked fundamental value. Casual inspection suggests that mispricing was even smaller than in CK, but an analysis of the bubble measures (defined below) showed that the difference was not significant under most specifications.

Figure 1: Median absolute deviation per treatment



Notes: The figure shows aggregated measures of mispricing by treatment. For each market and period, we calculate the absolute deviation of the median transaction price from fundamental value. The figure then shows the median over the six markets for each treatment in each period.

What is not seen in figure is that there is considerable heterogeneity in the performance of the six markets in each treatment. Appendix B shows the median transaction prices per period in each market for each treatment. Figure B1 shows the price trajectories for the six Baseline markets. Consistent with previous research on SSW-type markets, we observe a tendency for prices to start out below fundamental value in the first period. Two markets go on to exhibit the typical bubble-and-crash pattern (note that the use of median prices masks the crash in market 2 which had a closing price of 150), two markets remain flat for much of the experiment, while the remaining two markets largely track fundamental value. Overall, the Baseline markets are characterized by large price deviations from fundamental value and highly heterogeneous price trajectories, which is not uncommon in SSW markets.

In the CK markets (Figure B2) we again observe trade occurring at prices below fundamental value in the first period. However from period two onward, prices in four of the

six markets track very closely in line with fundamental value.[9] In the remaining two markets we observe a residual tendency for shares to trade above fundamental value through the second half of the experiment, but even this is mild compared to the mispricing typically observed in inexperienced SSW markets. In general, as shown in figure , the CK markets do not display the pronounced tendency to bubble and crash that is the norm with inexperienced subjects.

In sharp contrast to our CK markets, the absence of common knowledge under NCK is associated with large price deviations from fundamental value in several markets (Figure B3). In particular, we observe overpriced transactions[10] in three of the six markets. Overall, this treatment displays some of the most severe instances of mispricing that we observe in our entire study. The WAIT markets (Figure B4) appear similar to the Baseline in that we observe some markets in which there are price bubbles, some in which the price remains more or less flat over time, and one market that tracks fundamental value.

In short, while the six markets in each of the treatments are not identical in the extent of mispricing, figure does yield a fairly representative picture of the effect of our treatments. We now formalize the analysis of treatment effects by calculating standard bubble measures and using regression analysis.

4.2 Bubble measures and regression analysis

We follow the literature on asset market experiments in computing measures of the severity of mispricing. The literature has produced an abundance of different measures and there exists no golden standard as to which measures to present. Stöckl, Huber, and Kirchler (2010) discuss various of these bubble measures in the light of what they see as three essential criteria[11]. They conclude that the traditional bubble measures do not fulfill these criteria and instead suggest calculating two alternative bubble measures. The first measure, *Relative*

---

[9] In a seventh market, the prices tracked fundamental value from period two through to twelve, at which time we experienced a fatal server crash. The data from this crashed market are not included in the analysis reported below.

[10] We follow (Palan (2009)) in defining an overpriced transaction as one that occurs at a price in excess of the maximum dividend holding value.

[11] The criteria are that bubble measures should (i) relate price and fundamental value, (ii) be monotone in the difference between price and fundamental value, and (iii) be independent of the number of periods and the absolute level of the fundamental value.

Deviation (RD), is a measure of overpricing calculated as the average (over the T = 15 periods) difference between mean transaction price and fundamental value, normalized by average fundamental value $\left(\left|\bar{f}\right| = 192\right)$. Formally:

$$Relative\,Deviation = \frac{1}{T}\sum_{t=1}^{T}\left(\overline{P}_t - f_t\right)\Big/\left|\bar{f}\right|$$

where $\overline{P}_t$ is the mean transaction price in period t and f_t is the fundamental value in period t. Note that the definition of RD implies that positive and negative price deviations from fundamental value cancel out. A relative deviation of 0.3 means that the assets on average are overvalued by 30 percent compared to $\left|\bar{f}\right|$.

The second measure, *Relative Absolute Deviation* (RAD), is a measure of mispricing. It is very similar to RD but is calculated using the absolute (instead of the raw) difference between mean price and fundamental value. Formally, RAD is defined as:

$$Relative\,Absolute\,Deviation = \frac{1}{T}\sum_{t=1}^{T}\left|\overline{P}_t - f_t\right|\Big/\left|\bar{f}\right|$$

Thus, positive and negative price deviations do not cancel out in RAD. Intuitively, a relative absolute deviation of 0.3 means that mean prices on average differ 30 percent from the average fundamental value in the market, $\left|\bar{f}\right|$. Hence, for both measures, a larger value indicates more severe deviations of price from fundamental value.

The top panel of table 3 reports the mean values of the two measures for each of the four treatments. It shows that the RD yields negative values for three of the four treatments, reflecting the substantial underpricing occurring in several markets. In absolute terms, CK has the smallest RD (0.027), WAIT the second smallest (0.071) while NCK and Baseline have almost identical values (0.106 and 0.101, respectively). For RAD, CK again has the smallest mean value (0.182) while NCK has the second smallest (0.283) and Baseline and WAIT are very similar (0.370 and 0.348, respectively).

Table 3: Bubble Measure Analysis

	Relative Deviation	Relative Absolute Deviation
	Mean values	
Baseline	-0.101	0.370
CK	-0.027	0.182
NCK	0.106	0.283
WAIT	-0.071	0.348
	Wilcoxon rank-sum p-values	
Baseline vs. WAIT	0.749	0.873
CK vs. Baseline	0.337	0.078*
CK vs. Baseline + WAIT	0.512	0.039**
NCK vs. Baseline	0.200	0.631
NCK vs. Baseline + WAIT	0.160	0.512
CK vs. NCK	0.200	0.150
Common knowledge	0.947	0.039**
Control questions	0.204	0.094*

Notes: The top panel of the table shows the mean values of the two bubble measures (*Relative Deviation* and *Relative Absolute Deviation*) for the six markets in each of the four treatments. The bottom panel shows p-values for Wilcoxon rank-sum tests comparing the measures across treatments and groups of treatments. *Common knowledge* tests CK against the other three treatments; *Control questions* tests CK and NCK against Baseline and WAIT. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The lower panel of table 3 reports p-values for Wilcoxon rank-sum tests comparing the two measures across treatments. The first row confirms that Baseline and WAIT do not differ significantly from one another ($p > 0.749$) for either measure. The second and third rows show that CK differs significantly in RAD, but not in RD, from Baseline at the 10 percent level, and from Baseline pooled with WAIT at the 5 percent level. Rows four and five show that NCK is not significantly different from either Baseline or Baseline pooled with WAIT for either measure. Row six shows that CK and NCK are not different for either measure (p = 0.200 for RD and p = 0.150 for RAD). Row seven tests the effect of common knowledge, i.e. treatment CK vs. the other treatments. We find that the effect of common knowledge is significant for RAD (p = 0.039) but not for RD (p = 0.947). The last row of table 3 shows that control questions make no difference for RD but are weakly significant for RAD (p = 0.094).

To summarize, the analysis of bubble measures shows that *Relative Deviation* is not significantly different across treatments. This is because we observe both over- and underpricing which cancel out in the calculation of RD, and average overpricing is thus rather small in all treatments. This is not the case for *Relative Absolute Deviation* where we find a significant difference between CK and the markets without control questions (Baseline and WAIT). For RAD, we find that common knowledge significantly decreases mispricing (when compared to all treatments without common knowledge). We find a weaker effect of control questions in themselves when comparing treatments with and without the questions. This suggests that common knowledge may be more important than just control questions.[12]

The advantage of using measures of mispricing in combination with nonparametric statistical tests, as we do above, is that this is a very simple and straightforward approach which entails few assumptions. However it is also limited, in that the measures are highly aggregated and thus disregard a lot of information. In addition, the nonparametric tests do not allow us to make partial inference when pooling treatments into one dataset. For example, we are not able to distinguish between treatments CK and NCK when testing for the overall effect of control questions (in the last row in table 3), even though these two treatments are fundamentally different. This means that we cannot say whether the weakly significant effect is driven by the fact that it is common knowledge that everyone has to answer the control questions in CK, or by the control questions *per se*.

To take into account these drawbacks, we extend the analysis above using regression analysis. To do so, we first define two measures based on the median price in each period and market. In particular, we define $AbsDev_t$ to be the absolute difference between the median transaction price and the fundamental value in a period:

$$AbsDev_t = \left| \tilde{P}_t - f_t \right|$$

Furthermore, we create a binary variable $Track_t$ to indicate whether $AbsDev_t$ is within a tolerance band of 5% of fundamental value:

---

[12] We also calculate but do not present a number of the older bubble measures. The overall finding is that measures which focus on overpricing (*Amplitude*, as defined in King 1991, *Duration*, as defined in Porter and Smith 1995, *Average bias*, as defined in Haruvy and Noussair 2006) yield values that are similar across treatments while a measure that focuses on mispricing (*Total dispersion*, as defined in Haruvy and Noussair 2006) yields values that are smaller for CK than the other treatments.

$$Track_t = \begin{cases} 1 & \text{if } AbsDev_t < 0.05 \cdot f_t \\ 0 & \text{if } AbsDev_t \geq 0.05 \cdot f_t \end{cases}$$

We then define a set of dummy variables that characterize our different treatments. *CommonKnowledge* equals 1 in treatment CK; *ControlQuestions* equals 1 in treatments CK and NCK while *Time* equals 1 in all treatments except Baseline. The variable *Time* captures the fact that trading does not start immediately after the instructions have been read aloud.

We run OLS regressions using $AbsDev_t$ and $Track_t$ as dependent variables with standard errors clustered on individual markets to take account of correlated error terms[13]. The explanatory variables are the dummies described above (*ControlQuestions*, *CommonKnowledge* and *Time*) along with period dummies and a constant. Table 4 presents the results.

Table 4: Mispricing regressions

Dependent variable:	AbsDev	Track
ControlQuestions	-5.79	-0.100
	(17.01)	(0.101)
CommonKnowledge	-23.98 *	0.244 **
	(13.65)	(0.105)
Time	-4.53	0.077
	(20.65)	(0.113)
Constant	33.24 *	0.056
	(19.44)	(0.094)
Period dummies	Yes	Yes
Adj. R^2	0.214	0.085
N	359	359

Notes: The table shows OLS estimates (tobit regression for *AbsDev* censored at 0 and random effect regressions yield qualitatively similar results) on two measures of mispricing: *AbsDev* is the absolute deviation of period median prices from fundamental value. *Track* is a dummy variable equal to 1 if *AbsDev* is less than 5% of the fundamental value and zero otherwise. Parentheses show robust standard errors, clustered on markets. We exclude three outliers: one (0.015%) observation with price 2215 and two (0.030%) observations with price zero. * p<0.10, ** p<0.05, *** p<0.01

[13] Random effect regressions and tobit regressions censored at zero (for *AbsDev*) yield qualitatively similar results.

The first row of estimates in table 4 shows that control questions *per se* do not have a significant effect on the level of mispricing, for either of the two measures used.[14] However, the second row shows that common knowledge significantly reduces mispricing, when controlling for individual rationality (i.e. control questions). Thus, individual rationality alone does not affect mispricing but common knowledge of rationality does. This is in support of SSW's hypothesis that common knowledge of the fundamental dividend value is not necessarily enough to induce common knowledge of rationality and that this could explain (at least part of) the mispricing that we observe in SSW-type asset markets. Finally, we find that extra time for subjects to think about the situation does not affect mispricing.

5 Concluding remarks

When SSW first devised the design of their experiment, they intended it be a particularly simple and transparent bubble-free environment that would serve as a baseline for research into the factors that might contribute to the formation of price bubbles. Nonetheless, SSW were initially sanguine about their observation of price bubbles with inexperienced subjects, as they observed that rational expectations theory was not necessarily violated even if all traders were indeed rational but simply lacked common knowledge of this fact. This interpretation was consistent with the observation that bubbles were diminished with repeated experience as part of the same group of subjects. However, the results of Lei, Noussair, and Plott (2001) provided conclusive evidence of actual irrationality in inexperienced subjects.

As we interpret it, the Lei, Noussair, and Plott (2001) no speculation treatment eliminated the opportunity to speculate, yet did not deter subjects from making decision errors – indeed it was designed precisely to demonstrate that subjects were indeed making such errors. Our NCK treatment in some respects does the opposite. In particular, our control questions are intended to improve rationality at the individual level, and thus reduce the extent of decision errors. However, because the NCK treatment does not create common expectations, it gives subjects who do not believe in the rationality of others a motive to speculate. The result, as we have shown, is substantial mispricing. Hence we conclude that –

[14] The results are robust to using the mean instead of the median or of using the median of absolute prices instead of using absolute median prices for *AbsDev*, as well as to using different definitions of *Track*, including 10% deviation or an absolute deviation of 10 experimental currency units.

taken alone – neither removing the possibility of speculation nor reducing decision errors is sufficient to eliminate mispricing.

Incidentally, this interpretation aligns with that of Noussair and Plott (2008), who attribute the bubble and crash phenomenon to two sources. Firstly, even though all traders may be rational, if this rationality is not common knowledge then some may still hold the belief that there are irrational traders in the market, and this may lead to speculation that drives prices above fundamental value. Secondly, mispricing may also simply reflect actual decision errors on the part of some subjects.

Our CK treatment succeeds in largely eliminating mispricing because it addresses each of these points. Firstly, we require that subjects properly understand the decision environment in order to correctly answer our battery of control questions. This minimizes decision errors. Second, the fact that it is common knowledge that all traders must correctly answer all of these questions reduces uncertainty regarding the behavior of others. This creates common expectations and reduces the scope for speculation. The result is that mispricing is essentially eliminated in four of our six CK markets – even though our traders are inexperienced – with only small deviations from fundamental value in the remaining two markets. Our results thus speak directly to SSW's conjecture regarding the importance of homogeneous expectations in the formation of bubbles, by showing that common knowledge of rationality is sufficient to largely eliminate the bubble-and-crash phenomenon that has occupied researchers for over twenty years.

References

van Boening, Mark V., Arlington W. Williams, and Shawn LaMaster, 1993, Price bubbles and crashes in experimental call markets, *Economics Letters* 41: 179–185.

Fischbacher, Urs, 2007, z-Tree: Zurich toolbox for ready-made economic experiments, *Experimental Economics* 10: 171–178.

Haruvy, Ernan, and Charles N. Noussair, 2006, The Effect of Short Selling on Bubbles and Crashes in Experimental Spot Asset Markets, *Journal of Finance* 61: 1119–1157.

King, Ronald R., 1991, Private Information Acquisition in Experimental Markets Prone to Bubble and Crash, *Journal of Financial Research* 14: 197–206.

King, Ronald R., Vernon L. Smith, Arlington W. Williams, and Mark V. van Boening, 1993, The Robustness of Bubbles and Crashes in Experimental Stock Markets, in Richard Hollis Day, and Ping Chen, eds.: *Nonlinear dynamics and evolutionary economics* (Oxford University Press, New York).

Lei, Vivian, and Filip Vesely, 2009, Market Efficiency: Evidence from a No-Bubble Asset Market Experiment, *Pacific Economic Review* 14: 246–258.

Lei, Vivian, Charles N. Noussair, and Charles R. Plott, 2001, Nonspeculative Bubbles in Experimental Asset Markets: Lack of Common Knowledge of Rationality vs. Actual Irrationality, *Econometrica* 69: 831–859.

Noussair, Charles N., and Steven J. Tucker, 2006, Futures Markets and Bubble Formation in Experimental Asset Markets, *Pacific Economic Review* 11: 167–184.

Noussair, Charles N., and Charles R. Plott, 2008, Bubbles and Crashes in Experimental Asset Markets: Common Knowledge Failure?, in Charles Raymond Plott, and Vernon L. Smith, eds.: *Handbook of experimental economics results* (North Holland, Amsterdam).

Palan, Stefan, 2009, *Bubbles and crashes in experimental asset markets* (Springer, Heidelberg).

Porter, David P., and Vernon L. Smith, 2008, Price Bubbles, in Charles Raymond Plott, and Vernon L. Smith, eds.: *Handbook of experimental economics results* (North Holland, Amsterdam).

Porter, David P., and Vernon L. Smith, 1995, Futures Contracting and Dividend Uncertainty in Experimental Asset Markets, *Journal of Business* 68: 509–541.

Smith, Vernon L., 1994, Economics in the Laboratory, *Journal of Economic Perspectives* 8: 113–131.

Smith, Vernon L., Gerry L. Suchanek, and Arlington W. Williams, 1988, Bubbles, Crashes, and Endogenous Expectations in Experimental Spot Asset Markets, *Econometrica* 56: 1119–1151.

Smith, Vernon L., Mark V. van Boening, and Charissa P. Wellford, 2000, Dividend timing and behavior in laboratory asset markets, *Economic Theory* 16: 567–583.

Stöckl, Thomas, Jürgen Huber, and Michael Kirchler, 2010, Bubble measures in experimental asset markets, *Experimental Economics* 13: 284–298.

Appendices

Appendix A: Control questions

Appendix B: Median price trajectories in individual markets

Appendix C: Instructions

Appendix A: Control Questions

In total, we require subjects in the CK and NCK treatments to correctly answer 34 control questions: 4 in the basic set and 30 in the extended set. The extended set consists of two series of 15 questions: one series which is framed from the perspective of a buyer and one which is framed from the perspective of a seller. The questions are stated below.

Basic set (4 questions)

Question 1: What is the average dividend from the share in period 14?

Question 2: What is the total average dividend that you will receive if you hold the share from period 14 and to the end of the market (i.e. until period 15)?

Question 3: What is the total maximum dividend that you will receive if you hold the share from period 14 and to the end of the market (i.e. until period 15)?

Question 4: What is the total minimum dividend that you will receive if you hold the share from period 14 and to the end of the market (i.e. until period 15)?

Extended set of control questions, buyer frame (15 questions)

For $i = \{15, 14, \ldots, 1\}$, subjects are asked:

Question 5 + (15-i): Suppose that you buy one share in period i and that you keep it until the end of the market (i.e. until period 15). What is the average total dividend that you will receive from this share?

Extended set of control questions, seller frame (15 questions)

For $j = \{15, 14, \ldots, 1\}$, subjects are asked:

Question 20 + (15-j): Suppose that you sell one share in period j and that you do not buy it back. What is the average total dividend that you give up on this share?

Appendix B: Median price trajectories in individual markets

This Appendix shows the median transaction price trajectories in each of the six markets for each of our four treatments: Baseline (Figure B1), CK (Figure B2), NCK (Figure B3) and WAIT (Figure B4).

Figure B1: Median Transaction Prices, Baseline markets

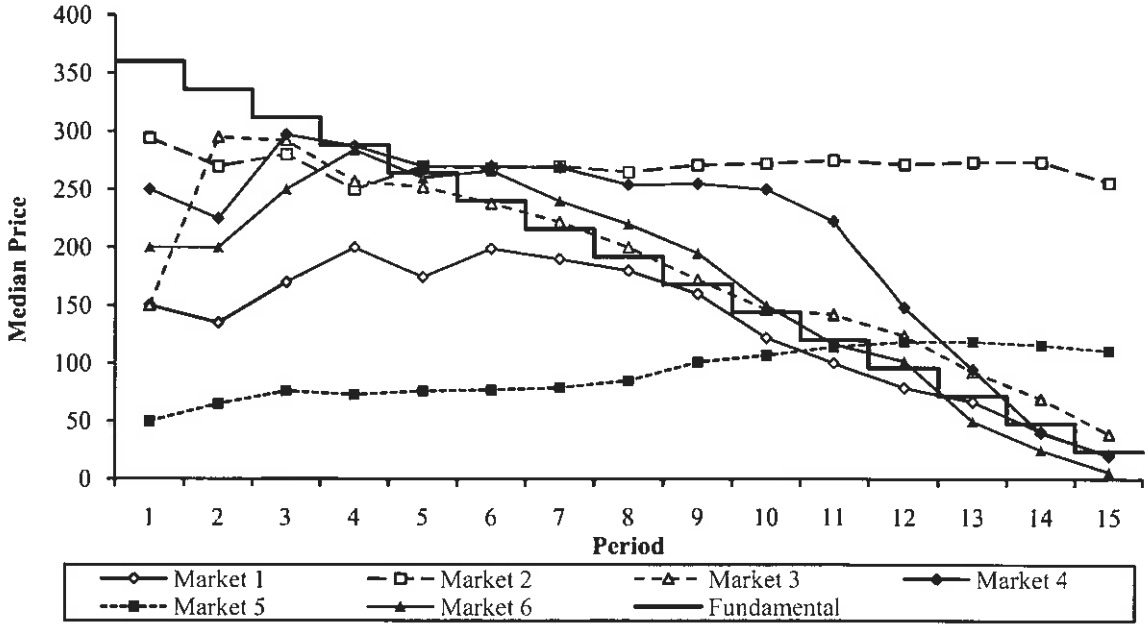


Figure B2: Median Transaction Prices, CK markets

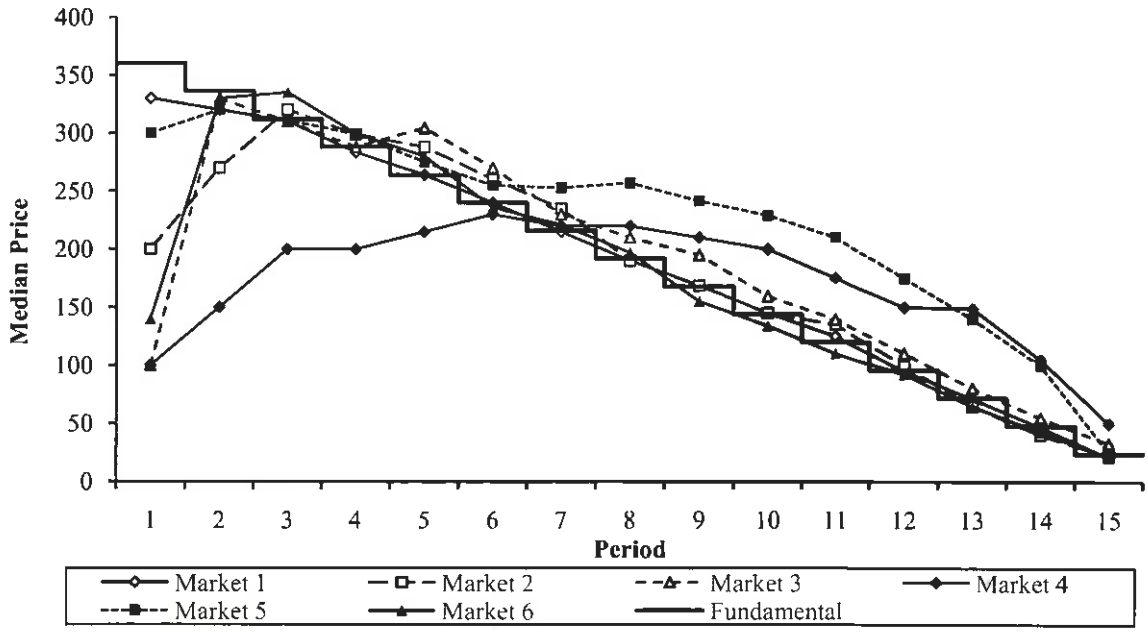


Figure B3: Median Transaction Prices, NCK markets

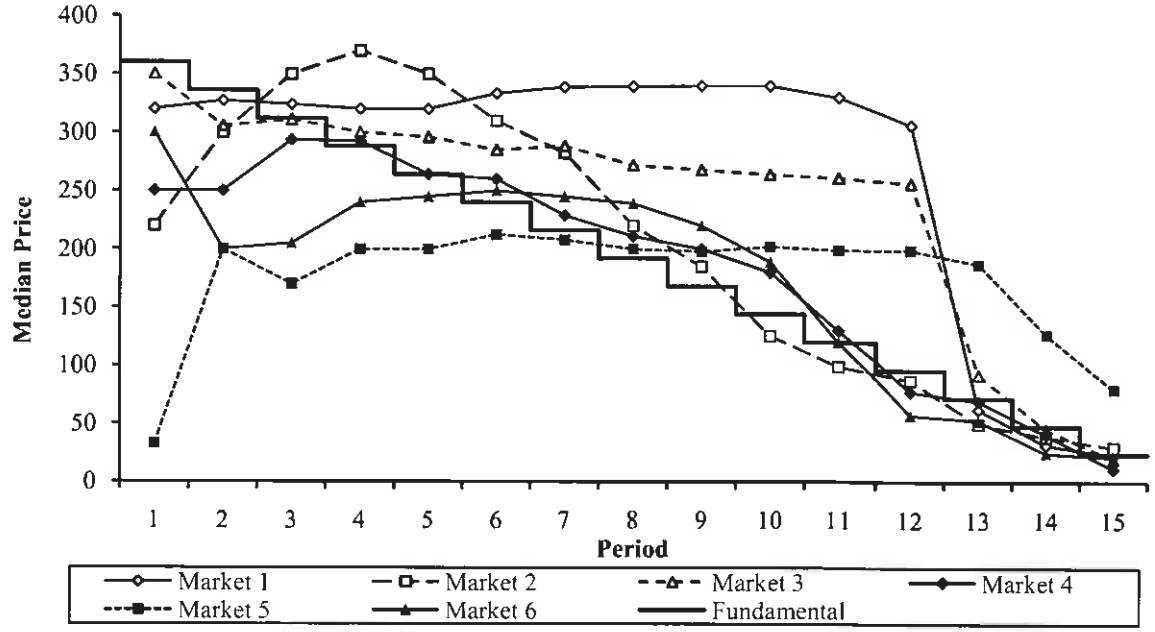
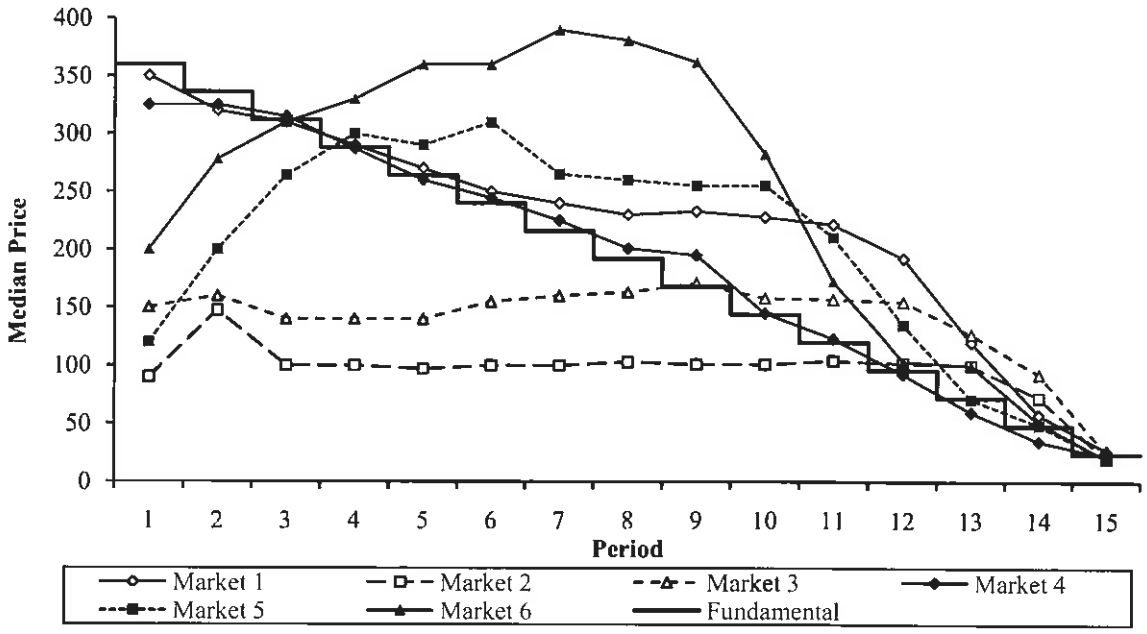


Figure B4: Median Transaction Prices, WAIT markets



Appendix C: Instructions for the CK treatment ‡

General Instructions

This is an experiment on decision making in a market. The instructions are simple and if you follow them carefully and make good decisions, you may earn a considerable amount of money which will be paid to you in cash at the end of the experiment.

Please do not communicate with other participants during the experiment. If you have a question please raise your hand, and an experimenter will assist you.

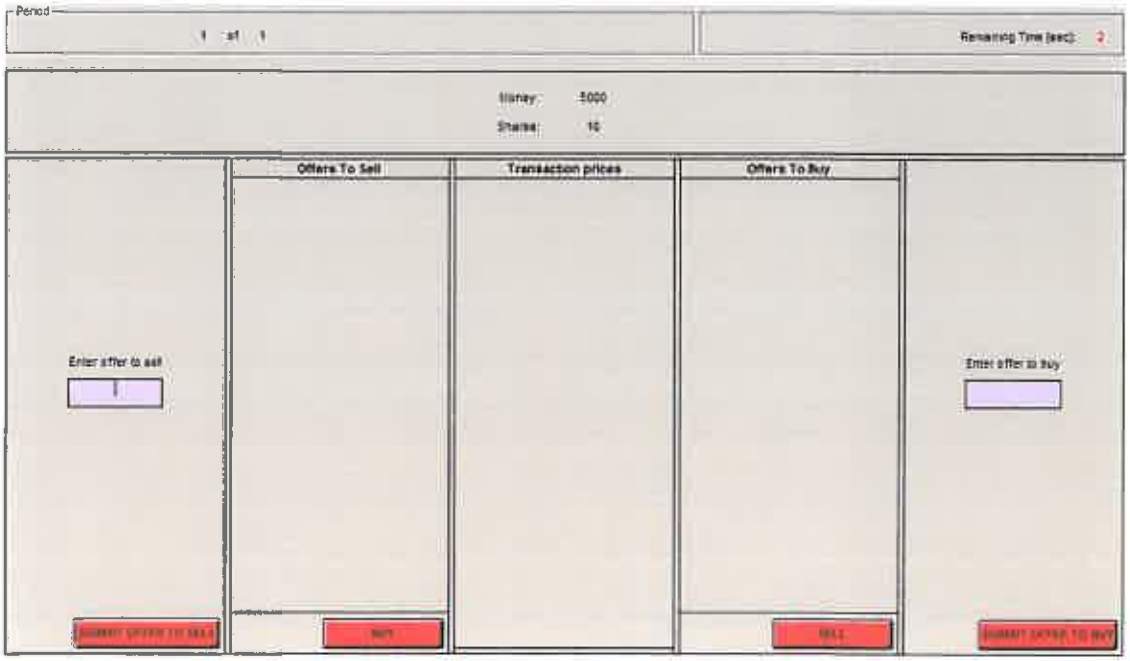
In this experiment, you have the opportunity to buy or sell in a market. The money used in this market is 'Experimental Currency Units' (ECU). All trading will be done in terms of ECU. The cash payment to you at the end of the experiment will be in Danish kroner.

The conversion rate will be **11 ECU to 1 krone.**

You will then be asked to complete a questionnaire, after which you will receive your payment. The entire experiment will last approximately two-and-a-half hours, including half an hour for instructions and practice.

How to use the Computerised Market

On the top right of the screen you will see how much time is left in the current trading period. The items you can buy and sell in the market are called shares. In the centre of your screen you will see the number of shares and the amount of money you currently have.



---

‡ Horizontal rules denote the positions of the page breaks in the original instructions.

The screen can be used to participate in the market in one of four ways.

Making an offer to sell a share, by entering the price at which you would like to sell:

To offer to sell a share, enter the price at which you would like to sell in the box labelled 'Enter offer to sell' on the left of the screen, then click on the button 'Submit offer to sell'.

The second column from left will show a list of offers to sell, each submitted by a different participant. The lowest offer-to-sell price will always be on the bottom of the list. Your own offer will appear in blue. Submitting a new offer will replace your previous one.

Making an offer to buy a share, by entering the price at which you would like to buy:

To offer to buy a share, enter the price at which you would like to buy in the box labelled 'Enter offer to buy' on the right of the screen, then click on the button 'Submit offer to buy'.

The second column from right will show a list of offers to buy, each submitted by a different participant. The highest offer-to-buy price will always be on the bottom of the list. Your own offer will appear in blue. Submitting a new offer will replace your previous one.

Buying a share, by accepting an offer to sell:

You can select an offer to sell in the second column from left by clicking on it. If you click the 'Buy' button at the bottom of this column, you will buy one share at the selected price. However you are not allowed to buy a share from yourself.

When you accept an offer to sell, it will disappear from the list. If you had also placed an offer to buy, it will disappear from the offers to buy list because you have just bought a share.

Selling a share, by accepting an offer to buy:

You can select an offer to buy in the second column from right by clicking on it. If you click the 'Sell' button at the bottom of this column, you will sell one share at the selected price. However you are not allowed to sell a share to yourself.

When you accept an offer to buy, it will disappear from the list. If you had also placed an offer to sell, it will disappear from the offers to sell list because you have just sold a share.

Transaction prices

When you buy a share your money decreases by the price of the purchase. You can only buy a share if you have enough money to pay for it.

When you sell a share your money increases by the price of the sale. You can only sell a share if you owned one to begin with.

In the middle column of the screen, labelled 'Transaction prices', you will see the prices at which shares have traded in the current period.

Practice period

You now have ten minutes to practice buying and selling shares. Your actions in this practice period will not influence your earnings or your position later in the experiment. The only goal is to master the use of the interface.

Please make sure that you successfully submit offers to buy and offers to sell. Also make sure that you successfully accept other people's offers to buy and sell shares.

If you have any questions, please raise your hand and an experimenter will assist you.

[DISTRIBUTED AFTER THE PRACTICE PERIOD]

Specific Instructions for this Experiment

In each market there are ten participants. *Although there may be more than ten participants in the lab today, you will always be in the same market of ten participants, consisting of yourself and the same set of nine others.*

The market will consist of fifteen trading periods. In each period there will be four minutes during which you can trade shares in exchange for ECU.

At the beginning of the first trading period, your screen will display your initial holdings of money and/or shares. These will not necessarily be the same for all participants in the market.

Any trade that you make will change your holdings of money and shares. These holdings will carry over from one trading period to the next.

Dividends

Recall that the market consists of fifteen trading periods. Shares are assets with a life of fifteen periods. Each share will pay a dividend to its current owner at the end of each period.

The dividend is randomly determined by the computer, and will be the same for all shares. In particular, each share that you own at the end of a period will pay:

- a dividend of 0 ECU with probability 1/4;
- a dividend of 8 ECU with probability 1/4;
- a dividend of 28 ECU with probability 1/4; and
- a dividend of 60 ECU with probability 1/4.

Since each outcome is equally likely, the average dividend is (0+8+28+60) / 4 = 24 ECU in every period.

Dividends will be added to your money balance automatically at the end of each period. After the dividend is paid at the end of the fifteenth trading period, all shares will be worthless and there will be no further earnings possible from them.

Average Holding Value Table

You can use your AVERAGE HOLDING VALUE TABLE to help you make decisions.

The first column indicates the Ending Period of the market. The second column indicates the Current Period for which the average holding value is being calculated. The third column gives the Number of Holding Periods from the Current Period to the Ending Period.

The fourth column gives the Average Dividend per Period for each share that you hold. The fifth column gives the Average Holding Value per Share that you hold from the Current Period until the end of the market.

That is, for each share that you hold for the remainder of the market, you will earn on average the amount listed in column five. The value in column five is calculated by multiplying the values in columns three and four.

AVERAGE HOLDING VALUE TABLE

Ending Period	Current Period	Number of Holding Periods	×	Average Dividend Per Period	=	Average Holding Value Per Share
15	1	15		24		360
15	2	14		24		336
15	3	13		24		312
15	4	12		24		288
15	5	11		24		264
15	6	10		24		240
15	7	9		24		216
15	8	8		24		192
15	9	7		24		168
15	10	6		24		144
15	11	5		24		120
15	12	4		24		96
15	13	3		24		72
15	14	2		24		48
15	15	1		24		24

Your Earnings

At the end of the market, your earnings will equal the amount of money you have at the end of period fifteen, after the last dividend has been paid.

This amount of money will be equal to:

Any money you had at the beginning of period one
+ Any money you received from sales of shares
− Any money you spent on purchases of shares
+ Any dividends you received

At the conclusion of the experiment this amount will be converted into Danish kroner at the rate specified on page one of these instructions, and paid to you in cash.